UNIVERSITY OF CALIFORNIA
Santa Barbara

Modular wireless networks for infrastructure-challenged environments

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Mariya Zheleva

Committee in Charge:

Professor Elizabeth Belding, Chair

Professor Heather Zheng

Professor Ben Zhao

Ranveer Chandra

September 2014

The dissertation of
Mariya Zheleva is approved:

Professor Heather Zheng

Professor Ben Zhao

Ranveer Chandra

Professor Elizabeth Belding, Committee Chairperson

July 2014

Modular wireless networks for infrastructure-challenged environments

Mariya Zheleva

To my best friend and partner in life Petko, my sister Dessy, my parents Elena and Zhivko and all the other amazing people who inspire me.

Acknowledgements

First and foremost I would like to thank my life partner Petko for standing next to me. Thank you for your endless love, for holding my hand through hard moments and for the inspiration!

I would also like to thank my research adviser Professor Elizabeth Belding. You are an amazing mentor and a great example for a teacher, researcher and person, and I am fortunate to have learned from you. Thank you for all the opportunities to explore myself as a researcher and the world in (and out) of the scope of our research projects.

I also thank my committee members Professor Heather Zheng, Professor Ben Zhao and Ranveer Chandra for the valuable feedback, great classes and fruitful collaborations.

This work would not have been so much fun without the help of all my collaborators: Veljko Pejovic, David Johnson, Paul Schmitt, Morgan Vigil, Arghyadip Paul, Danny Iland, Ashish Kapoor, Aakanksha Chowdhery, Kate Milosavljevic and the people from LinkNet at MachaWorks. I would like to extend a very special "thank you" to Veljko and David: your great example and dedication to ICTD is fundamental to my growth as a researcher in this field.

I would like to express my endless gratitude to my family and friends. Dessy, mom and dad, you have been a great support throughout the years and have helped develop my curiosity and dedication to teach, which I have found essential to completing this degree. Last but not least, I thank my friends Ivo, Ani, Zhivko, Petyo, Marta, Mariya, Danny, Dancho, Alex, Veljko, Carina, Martin, Chris, Lau, Cara, Beector, Eugeniu, Johannes, Jess, Paul, Shannon, Dimo and all the lovely people I know for helping me keep my balance and inspiring me every day. Much love to all.)◇(

Curriculum Vitæ

Mariya Zheleva

Education

2014	*Doctor of Philosophy (Ph.D.)* Dept. of Computer Science, University of California, Santa Barbara, California, USA.
2013	*Master of Science (M.S.)* Dept. of Computer Science, University of California, Santa Barbara, California, USA.
2008	*Master of Engineering (M.Eng.)* Dept. of Telecommunications, Technical University of Sofia, Sofia, Bulgaria
2006	*Bachelor of Engineering (B.Eng.)* Dept. of Telecommunications, Technical University of Sofia, Sofia, Bulgaria.

Experience

2009 – 2014	*Research/Teaching Assistant* Dept. of Computer Science, University of California, Santa Barbara, California, USA.
2013	*Research Intern* Microsoft Research, Redmond, WA, USA.

2011	*Research Intern* Broadcom Inc., San Diego, CA, USA.
2010	*Software Engineer Intern* RightScale, Santa Barbara, CA, USA.
2008 – 2009	*Data Network Support Engineer* Globul Ltd., Sofia, Bulgaria.
2006 – 2008	*Network Operation Center Engineer* Neterra Communications, Sofia, Bulgaria.
2003 – 2006	*Engineer* SG-Lab, Sofia, Bulgaria.

Awards

2013	*UCSB Graduate Division travel grant*
2010,2013	*NSF travel grant MobiSys10 and MobiSys13*
2012	*NSF travel grant DySpan12*
2011	*GHC11 Student Research Competition travel grant*
2009-2012	*Presidents Work-Study Award for 2009-2010, 2010-2011 and 2011-2012, UCSB*
2010,2013	*USENIX travel grant NSDI10 and NSDI13*

Publications

2014 **Mariya Zheleva**, A. Chowdhery, R. Chandra, A. Kapoor , P. Garnett, P. Mitchell, E. Belding, "TxMiner: Identifying Transmitters in Real-World Spectrum Measurements", *In Submission*.

2014 V. Pejovic, D. Johnson, **Mariya Zheleva**, A. Lysko, E. Belding, "VillageLink: Wide-Area Wireless Coverage", *in The Sixth International Conference on Communication Systems and Networks **(COMSNETS)***, Bangalore, India, January 2014.

2013 **Mariya Zheleva**, P. Schmitt, M. Vigil, E. Belding, "The Increased Bandwidth Fallacy: Performance and Usage in Rural Zambia", *in The Fourth Annual Symposium on Computing for Development **(ACM DEV)**, Cape Town, South Africa, December 2013*

2013 **Mariya Zheleva**, P. Schmitt, M. Vigil, E. Belding, "Bringing Visibility to Rural Users in Cote DIvoire", *n The International Conference on Information and Communication Technologies and Development **(ICTD)**, Cape Town, South Africa, December 2013*

2013 **Mariya Zheleva**, P. Schmitt, M. Vigil, E. Belding, "Community Detection in Cellular Network Traces", *in The International Conference on Information and Communication Technologies and Development **(ICTD)**, Cape Town, South Africa, December 2013.*

2013 P. Schmitt, M. Vigil, **Mariya Zheleva**, E. Belding, "Communication Flow Patterns in the Orange Telecom D4D Dataset", *in Third conference on the Analysis of Mobile Phone Datasets **(NetMob)**, Boston, MA, May 2013*

2013 **Mariya Zheleva** and M. Nekrasov, "Conferences: MobiSys 2013", *in **IEEE Pervasive Computing**, v. 13, p. 85. October-December 2013*

2013 **Mariya Zheleva**, A. Paul, D. Johnson, E. Belding, "Kwiizya: Local Cellular Network Services in Remote Areas", *in The 11th Interna-*

*tional Conference on Mobile Systems, Applications and Services (**MobiSys**), Taipei, Taiwan, June 2013*

2012 V. Pejovic, D. Johnson, **Mariya Zheleva**, E. Belding, L. Parks, G. van Stam, "The Bandwidth Divide: Obstacles to Efficient Broadband Adoption in Rural Sub-Saharan Africa", *in International Journal of Communications (**IJoC**), [S.l.], v. 6, p. 25, October 2012. ISSN 1932-8036*

2012 **Mariya Zheleva**, C. Budak, A. Paul, B. Liu, E. Belding, A. El Abbadi, "ImmuNet: Improved Immunization of Children Through Cellular Network Technology", 2nd best paper award *in UCSB Graduate Student Workshop in Computing (**GSWC**). Santa Barbara, CA, October, 2012*

Abstract

Modular wireless networks for infrastructure-challenged environments

Mariya Zheleva

While access to Internet and cellular connectivity is easily achieved in densely-populated areas, provisioning of communication services is much more challenging in remote rural areas. At the same time Internet access is of critical importance to residents of such rural communities. People's curiosity and realization of the opportunities provided by Internet and cellular access is the key ingredient to adoption. However, poor network performance can easily impede the process of adoption by discouraging people to access and use connectivity. With this in mind, we evaluate performance and adoption of various connectivity technologies in rural developing regions and identify avenues that need immediate attention to guarantee smoother technology adoption. In light of this analysis we propose novel system designs that meet these needs.

In this thesis we focus on *cellular* and *broadband Internet* connectivity. Commercial cellular networks are highly centralized, which requires costly backhaul. This, coupled with high price for equipment, maintenance and licensing renders cellular network access commercially-infeasible in rural areas. At the same time rural cellular communications are highly local: 70% of the rural-residential calls have an originator-destination pair within the same antenna [154]. In line with this observation we design a low-cost cellular network architecture dubbed Kwiizya [153], to provide local voice and text messaging services in a rural community. Where outbound connectivity is available, Kwiizya can provide global services. While commercial networks are becoming more available in rural areas they are often out of financial reach of rural residents. Furthermore, these networks typically provide only basic voice and SMS services and no

mobile data. To address these challenges, our proposed work allows Kwiizya to operate in coexistence with commercial cellular networks in order to extend local coverage and provide more advanced services that are not delivered by the commercial networks.

Internet connectivity in rural areas is typically provided through slow satellite links. The challenges in performance and adoption of such networks have been previously studied [50, 78]. We add a unique dataset and consequent analysis to this spectrum of work, which captures the upgrade of the gateway connectivity in the rural community of Macha, Zambia from a 256kbps satellite link to a more capable 2Mbps terrestrial link. We show that the improvement in performance and user experience is not necessarily proportional to the bandwidth increase [156]. While this increase improved the network usability, it also opened opportunities for adoption of more demanding services that were previously out of reach. As a result the network performance was severely degraded over the long term. To address these challenges we employ white space communication both for connectivity to more capable remote gateways, as well as for end user connectivity. We develop VillageLink [120], a distributed method that optimizes channel allocation to maximize throughput and enables both remote gateway access as well as end user coverage. While VillageLink features lightweight channel probing, we also consider external sources of channel availability [106]. We design a novel approach for estimation of channel occupancy called TxMiner, which is capable of extracting transmitter characteristics from raw spectrum measurements [152].

We study the adoption and implications of network connectivity in rural communities. In line with the results of our analyses we design and build system architectures that are geared to meet critical needs in these communities. While the focus of analysis in this thesis is on rural sub-Saharan Africa, the proposed designs and system implementations are more general and can serve in infrastructure-challenged communities across the world.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Access to connectivity has been growing across various domains. The number of cellular network subscriptions, for example, is approaching that of the world's population with over 120% coverage in developed and 89% coverage in developing countries. Further, as global optical networks become more available in the developing world [1] and Internet access becomes more affordable, the percentage of connected individuals and households increases as well. While these statistics are encouraging, in reality the majority of the population in developing regions currently remains disconnected; particularly disadvantaged are those living in rural areas. Our own experience in sub-Saharan Africa indicates that cellular network services are spotty and often of poor quality, forcing residents to walk kilometers before they can conduct a call, or carry multiple SIM cards and associate with different networks depending on availability [153].

Despite the perceived increase in Internet subscription, the fraction of residents of the developing world with access to fixed broadband is still as low as 6% and of those with mobile broadband – 20%. Further, the availability of these services is most often concentrated in urban areas and is typically out of financial and infrastructural reach to rural residents. In rural areas, mobile Internet access is provided through GPRS at best and fixed Internet access through slow and costly VSAT links. The low bandwidth of

GPRS services often fails to support basic Internet usage. Further, as a result of the high prices for VSAT connectivity, fixed Internet access is often a shared resource in rural communities, distributed among users in one of two ways. First established was the kiosk model [121] in which users go to a centralized location, e.g. Internet cafe, to get online. This access model results in restricted "deliberate [Internet] interaction" [148], subsequently leading to fundamentally limited usage. The second approach to shared access uses local Wi-Fi mesh networks to distribute the VSAT connectivity to residential [80, 96, 136]. While this allows for more leisure Internet access in comparison with the kiosk model, the major drawback of such approach is degraded user experience caused by the inherently poor performance of large-scale Wi-Fi mesh [38, 75, 131], which is further aggravated by the concurrent use of a limited connection by multiple users.

The implications of such poor network performance have been studied in recent work. Slow rural networks generally provide for low-bandwidth web-based applications and do not suffice for more bandwidth-intense usage such as video streaming, heavy downloads and upload of large files. Often even simple browsing activities fail due to high delays resulting in TCP timeouts, which eventually discourage users from retrying [50, 78]. Along with failure to access content, observed is a tendency in high failure rates in content upload [79]. These factors impede rural residents' abilities to access information online and have equal participation in Internet content generation. The latter, as recognized by van Hoorik [142], raises concerns about unreciprocal interaction of cultures threatening to stifle local traditions.

Traditionally, commercial network technologies have been designed to cover densely-populated areas with the goal of rapid return of investments. At the same time, rural areas are characterized with sparse populations and seasonal income. Thus, the cost and capacity at which rural networks are provisioned persistently lags behind those of

urban, which will keep taking its toll on network performance and user experience in the face of constantly-evolving and more bandwidth-hungry Internet applications [85]. In an effort to provide solutions, rural networks thus have been considered a "special kind of network" that require "custom solutions" [137]. Such solutions have been focused on either *network infrastructure* or *applications* design. One key aspect distilled from our analysis is that rural users have specific usage patterns, which opens space for custom designs catered to these patterns. In the context of cellular connectivity, we find that there is a high locality of interest in rural cellular communications whereby 70% of the rural-residential calls occur in the vicinity of the same cellular antenna [154]. In the context of Internet access too, Johnson et al. discover that the majority of Facebook interactions occur between local users [76]. Thus, technology for rural users needs to be designed with a twofold goal: first, to meet immediate needs for local interactions in the community and second, to be able to connect to global networks, where infrastructure is available, and dynamically adapt to the quality of outbound access.

1.1 Thesis Statement and Contributions

This dissertation shows that:

> *In order to improve user experience, rural networks and applications need to be highly modular and use this modularity to switch between local and global context according to network conditions and demand.*

In line with this statement, we design such modular systems to provide local cellular and Internet access in infrastructure-challenged communities.

The overview of this dissertation is depicted in Figure 1.1. We utilize a two-step approach in our solution design. We first employ sociological and technological methods in order to understand users' needs and network capabilities in the targeted communities. Based on our findings we then design solutions that meet actual user needs.

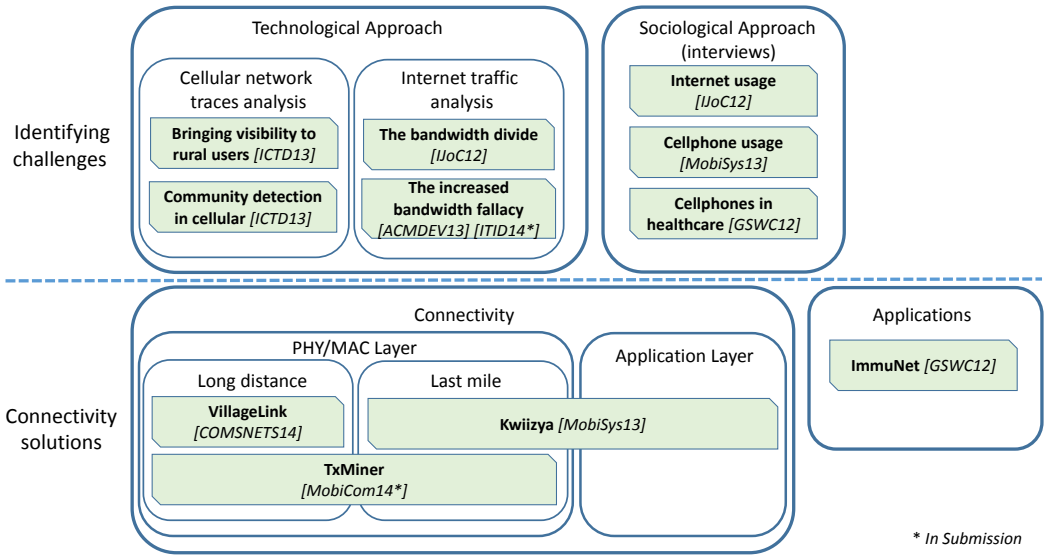


Figure 1.1: Dissertation overview. We design solutions after careful consideration of usage needs and abilities as well as social, regional and cultural factors. We study these factors in one of two ways. First, we rely on directly obtained information through face-to-face interviews. Second, we indirectly analyze usage patterns and technology actuation by analyzing network traces generated by the targeted communities.

Identifying challenges. We utilize two approaches to identify challenges in our partner communities: (i) sociological and (ii) technological.

- ***Technological approach.*** We base our technological approach on network traces analysis that aims to identify usage patterns and demand satisfaction. For this purpose we study both cellular as well as Internet network usage.

  Contributions: To determine cellular network use, we utilize a large-scale trace provided by Orange from their network in Cote D'Ivoire. We differentiate between rural and urban users in these traces and study network provisioning and utilization in rural areas. We find that rural users are largely under-provisioned in terms of cellular network infrastructure: while 98% of Cote D'Ivoire's territory is rural, only 43% of all antennas are in rural areas. We also find that there is

a *high locality of interest in rural cellular communications*, whereby 70% of all the rural-residential calls were conducted in the vicinity of the same antenna.

In order to understand Internet usage, we also study a network trace captured at the Internet gateway in the rural village of Macha, Zambia. We are particularly interested in the implications of an eight-fold bandwidth upgrade of the Internet link on user experience and network performance. We find that while usage patterns did not change immediately after the upgrade, services such as automatic software updates were more successful. As users began to harness the new opportunities for Internet experience, the usage focus shifted to content download, real time services (e.g. streaming and VoIP) and uploads. Consequently, this increased demand resulted in deteriorated network performance and user experience and as a result, by a few months after the upgrade, users became discouraged to attempt bandwidth-hungry applications such as peer-to-peer downloads and instead reverted to plain web browsing.

- ***Sociological approach.*** Our sociological approach is based on extensive in-person interviews with people from the targeted communities and is focused on cellphone use and applications. The goal of our interviews was two-fold: to understand cellphone usage, and to understand the feasibility of cellular network technology for health-care applications in the community. The interviews were conducted by the author with the help of our local partners who were in charge of recruiting interviewees and translating to English where necessary.

  Contributions: We find that while reasons for cellphone adoption are not drastically different than those indicated by Western users, the benefits to rural residents are much more pronounced. The poor communication between healthcare providers and patients arose as a desirable application of mobile telephony. Despite the obvious need and benefits, access to cellular services is still out of fi-

nancial reach for rural residents: few of the users could afford to buy talk-time on a regular basis and instead were using their phones as a receiving medium.

Connectivity solutions. We design communication solutions in line with the challenges identified in our analysis.

- ***Kwiizya.*** Mobile cellular technologies promise to bring connectivity to populations living in developing countries. At the same time such technologies are largely inaccessible to communities characterized by low income and sparse populations. The main reason for this unavailability are the high deployment and operational cost of commercial cellular networks with highly-centralized design.

  Contributions: To address the need of low-cost local cellular communication, we design Kwiizya, an open-source software and open-hardware based system that makes use of generic IP backbone to provide high-quality voice and text messaging in a rural community. We deploy an instance of Kwiizya in Macha, Zambia and evaluate feasibility and performance in an unregulated use-case. We find that Kwiizya is capable of delivering maximum MOS* across calls and has SMS delivery delay comparable to that provided by commercial networks†.

- ***VillageLink.*** While mobile cellular networks often are the shortest path to wide-spread connectivity, *computer-based* broadband access is still necessary for rural residents to fully embrace the opportunities provided by the Internet. Thus we endeavour to design solutions for wide-spread wireless broadband in rural areas. Our focus is on TV white spaces, which with their favourable propagation properties, are a great candidate for wide-spread broadband network coverage.

*Mean Opinion Score (MOS) is a metric that indicates perceived call quality. It is dependent on the voice codec and packet loss and varies between 1 and 5; the higher the MOS the better the call quality. Kwiizya's codec provides for a max MOS of 3.46; this is the value we observe across calls.

†The average time for SMS delivery in Kwiizya is 5.5 seconds and the maximum – 11 seconds. In comparison, measurements in Tanzania indicate that it takes between 6 and 16 seconds for text message delivery depending on provider.

Contributions: To this end we design VillageLink, a combined PHY/MAC mechanism for efficient channel allocation in Dynamic Spectrum Access (DSA) systems. VillageLink is designed to operate in a wide frequency band (50MHz to 800MHz) and makes use of existing TV antennas to connect end users. Operation in such a wide band poses some unique challenges in channel allocation, since the selection of a transmission frequency can impact the existence of a link. To address these challenges we make use of a two-step lightweight channel probing technique to understand the channel conditions at each communicating node. This creates a large solution space that calls for efficient methods to find the optimal channel allocation scheme. We design one such scheme that guarantees real-time convergence and identifies the optimal channel allocation to minimize interference while maximizing throughput. We show that our frequency-aware channel allocation leads to up to twice as much network capacity than an alternative heuristic based on interference avoidance.

- ***TxMiner.*** Traditionally, practical DSA systems have focused on TV white spaces as a medium. The latter have fairly static utilization in time and frequency because TV channels operate at a fixed bandwidth and either broadcast continuously or are silent. This makes the problem of scanning and frequency allocation fairly straightforward. To enable true opportunity for DSA technology, we consider DSA operations in frequency ranges beyond TV white spaces. In this context the problem of operation frequency selection becomes more complicated. Beyond TV white spaces, transmitters have much more dynamic characteristics, e.g. frequency-hopping (Bluetooth), aperiodicity (satellite transmissions) or varying channel width (802.11n/ac). To address this problem we design TxMiner: a system that *identifies transmitters from raw spectrum measurements without prior knowledge of transmitter signatures*.

Contributions: TxMiner harnesses the observation that wireless signal fading follows a log-normal distribution. Thus most of the signals emitted by the same transmitter will fall under a Gaussian curve in the dB scale. Harnessing this observation we develop a machine learning-based algorithm that uses Gaussian Mixture Models with Belief Propagation and can tease apart transmitters from spectrum measurements. We are particularly interested in features such as bandwidth, time or frequency duplexing, number of transmitters in a given bandwidth, mobility and directionality. We evaluate our algorithms on real spectrum measurements collected by Microsoft's Spectrum Observatory[‡], and show that TxMiner is able to identify various types of transmitters, with different physical layer protocols, in different parts of the spectrum.

- ***ImmuNet.*** Our interview analysis and informal conversations with health workers identify the poor information channels between patients and health-care providers as an important problem in health-care delivery in rural communities. To address this problem we propose a solution that is focused on improved delivery of vaccinations to children. Our solution is based on plain text messaging and provides capability for rapid outreach to large populations of patients as well as patient feedback to health workers.

  Contributions: We implement a prototype of Immunet that consists of *VaccStore*, a database that stores personal data, and *Kwiizya*. To enable rapid outreach, we enhance Kwiizya with an Instant Message to Text Message (IM-to-SMS) functionality that makes use of SMS broadcast, multicast and unicast for patient communication.

[‡]http://observatory.microsoftspectrum.com/

1.2 Dissertation Outline

The remainder of this dissertation is organized as follows. In Chapter 2 we present the background of our research including description of our field work as well as challenges in deploying rural networks. In Chapters 3 and 4 we analyze the performance and adoption of rural cellular and Internet networks. In Chapter 5 we present results from our social survey analysis. Our analysis indicates that there are specific usage patterns and needs in rural networks, which we address through the systems presented in Chapters 6, 7, 8 and 9. In Chapter 6 we present Kwiizya, a system that provides low-cost local cellular network services in remote areas. For scalability over a larger area, Kwiizya can make use of a solution such as VillageLink, presented in Chapter 7, in order to connect individual installations. While VillageLink features a light-weight spectrum-sensing technique, a comprehensive spectrum sensing approach is necessary in order to enable dynamic spectrum access in wide frequency range. To this end, we present TxMiner in Chapter 8. Finally, in Chapter 9 we describe an application that makes use of rural cellular networks such as Kwiizya for improved distribution of vaccines. We provide a discussion of future work and conclude this dissertation in Chapter 10.

Chapter 2

Research Background

The past decade has seen an increase in the number of mobile and fixed Internet subscriptions in developing countries. While these statistics are encouraging, what they mask is the huge discrepancy in access availability and affordability across regions of the world and between rural and urban areas within countries. 2014 statistics published by the International Telecommunications Union reveal that while 27% of the developed world has access to wired (or fixed) broadband, that number in developing countries is as low as 6%. Similarly, the mobile broadband subscriptions in developing countries are 21% in comparison with 84% in developed. The reason for such a stark difference in connectivity adoption is twofold: (i) the high discrepancy of Internet availability between rural and urban areas and (ii) the high cost of Internet access: while people in North America and Europe pay 1.8% of their monthly income towards connectivity, that number in Africa is as high as 56.5%.

The focus of our work is in Africa, where the rapid increase in submarine cables capacity terminating along the coast has brought hopes for price reduction. Unfortunately, the price of connectivity has remained unchanged. The main reason for this prices persistence is the lack of interconnecting infrastructure between the submarine optical cables terminating along the coast of Africa and the Internet exchange points on one hand, and the lack of national backbones and last mile infrastructure to connect end

users. Several factors dictate the lack of internal infrastructure in the African continent [18] including the low and non-uniform investments in the telecommunications sector, and the poor market conditions.

The competition introduced in the telecommunications sector in the 1990s along with the necessary transformations of the traditional telecommunication companies, from postal and telecommunications to GSM and shortly after that, CDMA, have required large investments by the incumbent operators in relatively dense intervals of time. In order to assure return of investments communication services became expensive immediately after these modifications. Unfortunately, despite the rapid adoption of the mobile network services, the end user prices for connectivity remain high to date. Several factors contributed to these high prices: first, with low-income populations the return of investments is slower; second, the international investments in telecommunications in Africa as compared to the rest of the world is low (e.g. 2.1% in 2004); and lastly, the distribution of investments is highly non-uniform across the African continent favouring the North African countries as well as South Africa, while the rest of African countries lack subsidies [18]. As a result of the high prices Internet connectivity is often of very limited capacity and is typically a shared resource to facilitate cost distribution among multiple stakeholders.

Another main reason for high connectivity prices is the poor market conditions in Africa, dictated by the lack of strong ICT industry along with the limited market size. The limited ICT production in Africa requires that all attributes for building and using communication networks, including production hardware and software as well as end user equipment, be imported from abroad. This further increases prices by adding transportation and customs cost. Lastly, the life-cycle of a western product is often not well-suited for the needs and financial capabilities of an African consumer: while westerners can afford to buy new software and hardware platforms as they are released

every 12-18 months, this is often not the case with the African consumer. Thus, the end user often makes use of outdated software and hardware that fails to meet the ever-increasing demand and capabilities of the modern Internet.

This dissertation assumes the context outlined above and studies the usage and performance of various networking solutions in rural sub-Saharan Africa. We study Internet access based on a longitudinal network trace collected at the Internet gateway in the rural village of Macha, Zambia between January 2011 and October 2012. We also use a large-scale cellular network trace spanning five months and provided by Orange Telecom to look at cellular network usage and provisioning in Cotê d'Ivoire. We identify unique characteristics of network usage and problems related to user experience and design systems to address these problems. In the remainder of this chapter we give background of our fieldwork and a description of the characteristics and challenges in working in the rural developing context.

2.1 Fieldwork

Our fieldwork focuses on the rural community of Macha, Zambia. Macha, highlighted in Figure 2.1, is a typical poor rural village located in the Southern province of Zambia. The village has a concentrated central area and a very dispersed population of 135,000 people, spread over a large radius of 35 km with average population density of 25 persons/km^2. Clusters of homes house a single extended family and the distance between separate households can be as much as several kilometres consisting of farmlands. The primary occupation in the village is maize farming. The average estimated income is $1/person/day – 5 times less than the round-trip cost to the closest town and 30 times less than a monthly Internet subscription limited to 1GB.

Figure 2.1: A map of Southern Africa highlighting the location of Macha.

Macha has been a local leader in health-care and technological innovation. Active organizations in the village include a hospital and health-care research facility as well as MachaWorks, an NGO that maintained a local wireless network, LinkNet*. LinkNet distributes Internet access from an Internet gateway over an area of 6 km^2, including schools, the hospital, the research institute and residential areas. LITA (LinkNet Information Technology Academy), an IT school affiliated with MachaWorks, teaches basic computer skills to local residents.

Internet usage and provisioning. While Macha is connected to the national power grid, electricity is rarely available in individual households. The lack of electricity coupled with the high prices for user equipment and Internet provisioning makes it virtually impossible for Machans to use Internet at home. Internet users in Macha typically access the Internet from work, from an Internet café or at school.

*The LinkNet organization existed during the time of this work but ceased operations in 2013. Currently, Internet services in Macha are provided through two of the national mobile operators – MTN and AirTel.

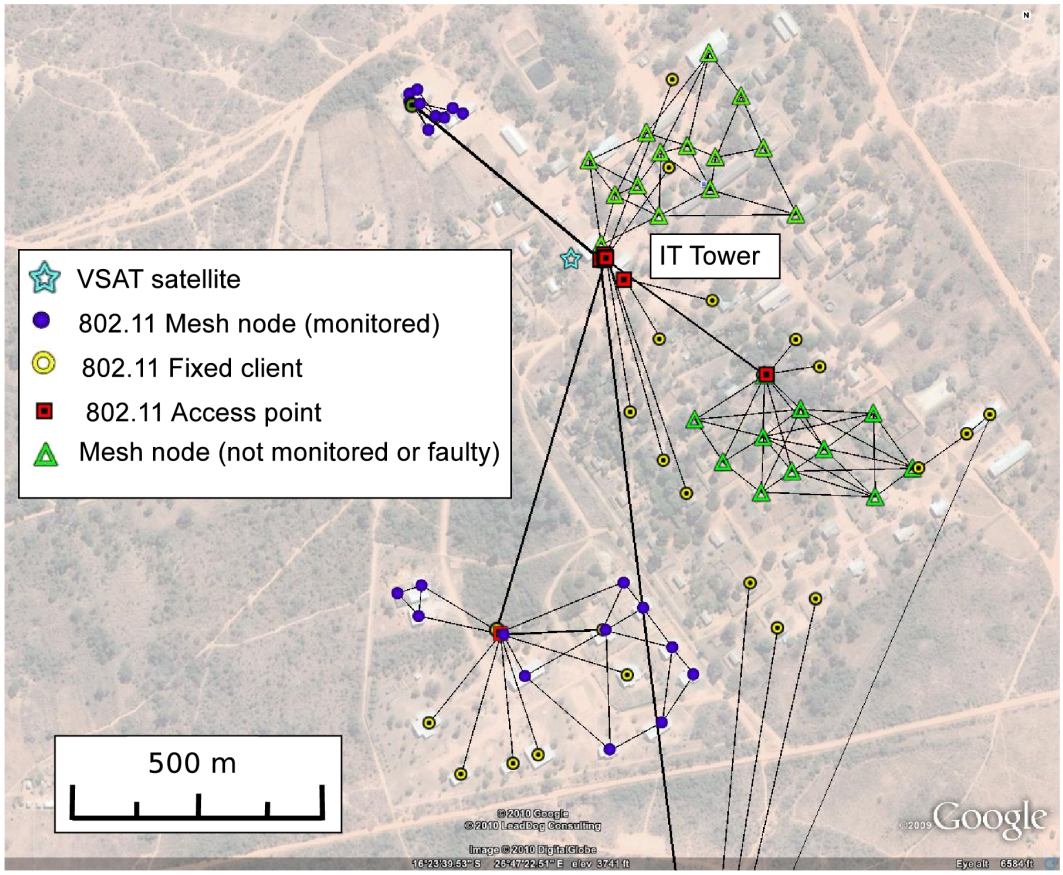


Figure 2.2: Wireless mesh network in Macha.

Internet access is distributed from the village Internet gateway to central facilities via a local wireless mesh network (Figure 2.2) maintained by LinkNet. We refer the interested reader to [96] for more details about the wireless mesh in Macha. Between 2008 and April 2011, the village was connected to the Internet through a satellite connection that cost $1200/month and provided 256kbps downlink bursting to 1Mbps, and 64kbps uplink bursting to 256kbps. In April 2011, the village Internet access was upgraded to a higher quality microwave terrestrial link with speeds up to 2Mbps costing $3600/month. At the time of the Internet link upgrade, approximately 300 residents were regular users of the Internet connectivity.

Cellular network access. Cellphone coverage in Macha was first introduced in 2006 by Celtel (now Airtel). By 2012, MTN was the second active cellphone provider in the village. Residents with a cellphone subscription can use plain voice and text messaging and, where available, low data rate GPRS service for Internet access. The coverage provided by the two operators is largely available in the central part of the village; coverage is inconsistent and spotty in residential areas. Residents often have SIM cards with both cellular providers to increase the likelihood they have cellphone coverage at any given time. Our first hand and anecdotal experience, however, indicates that the failures of both commercial networks are highly correlated and often coincide with power failures.

2.2 Rural Network Challenges

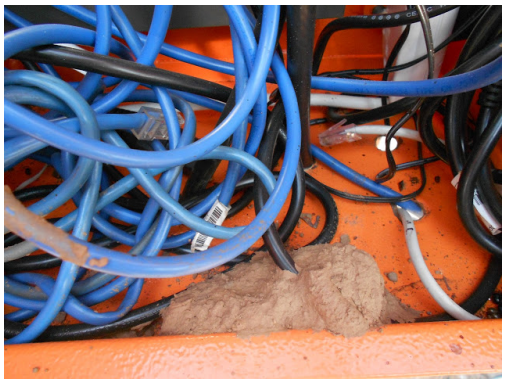
This section presents our experience in analysing Internet usage and deploying communication networks in Macha. This summary is based on traffic analysis, quantitative interviews conducted in the community as well as our anecdotal experience while spending time in the community.

2.2.1 Environment and Infrastructure

The unique landscape and shortage of infrastructure in the rural developing context poses challenges in the deployment and maintenance of communication networks. The import of equipment in a developing country can be challenging. Our anecdotal experience indicates that the best approach is to either ship equipment to an institution in a main city (e.g. a research institute or a university) or carry it as a check-in luggage upon arrival. In the latter case, the author's experience points that it is always safer to

(a) The main road in Macha

(b) The airstrip in Macha



(c) Wasp nest in a communication box.

(d) Harsh weather conditions in rainy season.

Figure 2.3: Environment and Infrastructure.

fly directly into the country of fieldwork; in the opposite case, additional inspections and/or corrupt activity can cause equipment to be detrimentally delayed or lost.

Once imported in the country, transportation of the equipment to the corresponding rural area is also a challenge. Naturally, with lacking transportation infrastructure in rural areas courier services are out of reach. Access to Macha is achieved either via a dirt road (Figure 2.3(a)) or through an unpaved airstrip that can handle landing of a small 4-seater air-plane (Figure 2.3(b)). Delivering equipment through these corridors is either cost-inefficient or hazardous to equipment. A one-way charter flight from the capital Lusaka to Macha cost $650 in 2013, which, given the small loads that these

plains can handle, adds a huge overhead in equipment transportation. Delivery through dirt roads, on the other hand, imposes hazards to communication equipment: dust entering the devices along with vibrations during transportation can cause equipment to fail even before reaching its destination.

Other infrastructure or environmental factors such as poor quality of electricity, harsh weather conditions and local habitat can pose serious challenges in the long-term operation and maintenance of communication equipment once it is deployed in a community. "Dirty" power can cause frequent and continuous interrupts of communication services. Persistent voltage instabilities can either damage equipment or boot it into a dysfunctional state. We provide a more in-depth analysis of power quality and its implications on Internet access in Chapter 6. Along with poor power, weather conditions and the local habitat can impose further challenges in network maintenance and performance (Figure 2.3(c) and 2.3(d)). For example, one of the cellular network base stations we deployed in Macha was damaged by lightning and another communication box became infested with wasps, imposing a threat of circuit short-cuts. Environment-proofing enclosures are essential for outdoor deployments in such conditions. Further, use of lightning arresters, reduction of exposed ethernet cables and predominant use of wireless links helps reduce the risk of lightning-related equipment failures.

2.2.2 Cultural

A variety of cultural characteristics can also pose challenges when rolling-out ICT projects in developing regions. The scheduling of day-to-day routines moves at a completely different pace than in many western countries. Plans are typically made in very short time scales, which can make collaboration hard and can take substantial adjustment for western researchers.

The common perception of westerners as a source of resources brings another challenge in working with under-serviced communities, which has previously been described as the researcher-pleasing phenomenon. Specifically, local residents or partners report what they think is the answer that will please the researcher, rather than the correct and actual answer. In our work we encountered a situation where residents continually reported that one of the systems we deployed was functional and they used it. It was not until we started analysing traces that we realized that the system was actually experiencing a problem that severely hampered performance.

Many African communities, including that of Macha, practice *ubuntu* traditions. Such cultures achieve peaceful coexistence through progress as a community rather than on an individual level. Innovations, including introduction of ICT, can substantially change the dynamics in such a community. To prevent such change in dynamics, researchers need to make sure they familiarize themselves with local dynamics and do their best to introduce a technology in a way that benefits the community as a whole and does not favour individuals or subgroups.

2.2.3 Skills and Local Expertise

The local ICT expertise in sub-Saharan African countries is very limited. For example, reports from Zambia and Tanzania [61, 95] indicate that only a few hundred people graduate with an ICT degree per year. As a result, the expertise necessary for ICT growth in rural sub-Saharan Africa is often external and temporally available. In the case of research projects in ICT, for example, expert researchers are typically on the ground only for the pilot deployment stage. Once researchers are back to their home institutions, pilot projects are often doomed to failure due to lack of local experts to facilitate maintenance and troubleshooting [31, 137]. That is why the involvement and training of local champions is very important for the long-term success of ICT projects.

There is a real challenge, however, in retaining locally trained experts. As described in [31], many organizations face a high IT staff turnover, whereby local people trained to support ICT projects take off to find a better-paid job in a larger city. Our deployments too have had similar fate. Several of our IT-trained local champions have moved to the Zambian capital for better jobs. Another partner remained in the village but has recently redirected all his energy into a more profitable business – in this case running a grocery store.

Lastly, when rolling-out an ICT project in a partner institution, researchers need to carefully evaluate the overhead this project will impose onto the local personnel. While it is often assumed that introducing a new technology would immediately alleviate the burden of a certain day-to-day routine, this might not always be the case. In Macha, for example, we worked with the primary healthcare division of the hospital on a project to support tracking of immunizations and provider-patient communication. While the staff members were enthusiastic about the system, it became clear that while learning to operate the new system they would need to retain their old practices as well. The latter was required by the ministry of health for reporting purposes. Thus, in such cases, while technology can give hope for simpler routines over the long term, immediately after deployment it might result in doubling the burden of already overworked staff, which subsequently may hamper the adoption of novel systems.

2.2.4 Economic

As is typical for many rural agricultural communities, the majority of people in Macha have seasonal income and rely on one-time sales of their produce (most often maize) in order to sustain their families throughout the year. As a result of this seasonality of income the affordability of connectivity varies largely throughout the year. While mobile cellular providers address this problem by offering pre-paid plans in very

small bundles, Internet service providers operate under contracts that require monthly bill payments. The latter makes it virtually impossible for individuals to get Internet connections.

The majority of Machans stay connected only through cellular operators, as opposed to Internet Service Providers, and can afford user equipment and services that support plain voice and text messaging. Mobile data is typically out of financial reach for two reasons: first, users cannot afford to buy smartphones and second, data bundles are too expensive. Furthermore, a majority of the residents can only afford to buy very minimal bundles – as much as to conduct a single call or to send several text messages, and instead carry their phones as a receiving media and wait for others to call them.

The cost of Internet access in a land-locked, rural community such as Macha is very high. The monthly fee for a satellite connection in Macha was $1200 and that of a 2MBps microwave terrestrial link was $3600. The high price and regularly-recurring bills makes it virtually impossible for individuals in the community to use Internet services. Thus, when Internet was first introduced in Macha it was a communal resource, managed and maintained by an organization called MachaWorks through their LinkNet project. The Internet connection was subsidized through international funding. In an effort to achieve self-sustainability the organization maintained an Internet café and a wireless mesh network, and sold individual access vouchers inspired by the pre-paid concept of mobile operators. A single voucher costed $30 – 30 times more than the average daily income, and was limited at 1GB data. This sustainability model, however, never broke even with the expenses for Internet connection and network maintenance. The project ceased to exist in late 2012 due to lack of external funding along with several stake-holders switching to data services available through one of the mobile operators.

2.2.5 Regulatory

Licensing poses a real challenge to deployment of novel wireless network technologies. In many developing countries spectrum regulations are unclear or lack all together. For example, the 2.4/5GHz Wi-Fi band is licensed for outdoors operation in some countries. Other countries such as the Philippines, lack licensing policies in some spectrum ranges and often have no clear information about what frequencies are being used and by what technologies. One of the collaborators of the author has been approached by the Philippines' government with request for a tool that can analyse spectrum occupancy and help identify incumbents. As indicated by the government representative, "Such a tool would enable us to better understand spectrum utilization and design our new spectrum policy".

When rolling-out a proof-of-concept deployment that operates in licensed spectrum, researchers often make use of experimental licenses. Beyond pilots, community mobile networks such as Rhizomatica in Mexico[†] also run under experimental licenses. While the procedure for obtaining such licenses in the US is well-documented, licensing for experimental purposes is widely unregulated in sub-Saharan Africa. In Zambia, there is a national regulatory organization that governs the distribution of spectrum resources; however, there is no policy for granting experimental licenses [105]. Furthermore, pricing of commercial licenses is oblivious to the potential return of investment. We argue that lower pricing or subsidizing commercial licenses for economically unattractive areas such as rural areas would encourage entrepreneurs to seek alternative technologies and deploy in such areas. Thus, while alternative technologies such as Kwiizya (presented later in this dissertation) have great technological potential to provide low-cost connectivity, the full potential of such systems is still dependent on favourable license policies and regulations.

[†]http://rhizomatica.org/

2.2.6 Technology

The high price of Internet connectivity in rural land-locked areas render Internet access practices that are very different than those in the western world. Westerners have high-speed access and can go online casually from home or the office. In contrast, in rural areas Internet is often a shared resource and is accessed from communal spaces such as Internet cafés and offices. Thus rural dweller's Internet access is much less casual, whereby they need to plan in advance what they need to do in the limited online time, prepare their e-mail correspondence offline and strictly synchronize with other parties in cases of real time communication (e.g. IM or Skype call) [78].

As many people need to share a limited Internet connection, rural networks typically suffer poor performance [78, 79]. As a result rural users have much more limited capabilities in exploring the Internet; our studies of user activity in Macha indicates that people most often used the connection to access basic websites, while activities such as video-streaming, real-time audio and content uploads failed. In a study of the Internet access upgrade, presented in Chapter 4, we unveil the desire of rural residents to fully utilize the services available online: following an eight-fold upgrade of the access connectivity users quickly began to utilize the link for more bandwidth-heavy applications.

2.3 Discussion and Conclusion

Rural areas pose unique challenges in the deployment of ICT projects. At the same time the benefits of adoption of ICT in rural communities is very clear. For example, access to Internet enabled rural farmers to adopt novel farming techniques, access to cellphones eliminated fraud in crops trading [153], and created novel ways of banking, which reduced theft where bank services were previously unavailable [98]. In

order for ICT projects to be truly beneficial in rural communities, they need to address actual communal needs and have to be supported from within the community. Several ICT projects have had long-term success. Such successful projects are managed in one of two ways: either by external experts who choose to stay long-term in the community [93, 128], or by strong local champions who take over after the initial deployment [66]. Moving forward, successful ICT initiatives should start from within the rural communities themselves. To facilitate such local initiatives, there is a need of a platform that enables local champions who are already involved in such projects to connect, work together, inspire each other and serve as role models to the upcoming generations of ICT experts.

This chapter discussed the complex combination of factors that govern the deployment and adoption of ICT projects in rural Africa. Projects that last beyond the test phase should utilize a holistic approach that strives to (i) study the specific needs and characteristics of a community where the project will be rolled out, (ii) design solutions that are carefully catered to community needs, and (iii) include local champions in system deployment and maintenance.

Our work strives to follow this approach. Through our collaboration with the rural community of Macha in Zambia as well as through network traffic analysis from Macha and Coté d'Ivoire we identify unique challenges in ICT work. Such challenges include environment and infrastructure, local culture, skills and local expertise, economics and regulations as well as technology.

Part I

Identifying Challenges in Developing Regions

The quality of provided communication services is essential to the successful adoption of these services. We begin with performance evaluation of various technologies for connectivity in developing rural regions and identify opportunities for improvement.

In many developing regions of the world, cellular technology is the first to provide access to telecommunications. Studies on adoption and usage of cellular networks have brought valuable insights into user-technology interaction. These studies serve as a base knowledge for development of critical services catered to specific user needs. We analyse cellphone network traces provided by Orange telecom from their network in Ivory Coast. First, we study emerging communicating patterns correlated with population density. We then examine community persistence in egocentric social graphs. We find that urban areas, although geographically distant, are strongly connected in terms of cellphone network activity. At the same time, cellular communication within rural residential areas is strictly local, whereby 70% of the calls in rural areas initiate and terminate at the same antenna. This drastic difference in urban and rural usage patterns warrants a rethink of conventional cellular network architectures and indicates that local solutions can be a feasible alternative for rural voice services, where the majority of calls occur between parties in close physical proximity.

Broadband Internet access too has become a critical part of socio-economic prosperity; however, only 6 in 100 inhabitants have access to broadband in developing countries. This limited access is driven predominately by subscriptions in urban areas. In rural developing communities, access is often provided through slow satellite, or other low-bandwidth long-distance wireless links, if available at all. As a result, the quality of the Internet access is often poor and at times unusable and there is a constant desire in users for increased access speeds. We study the performance and usage implications of an Internet access upgrade, from a 256kbps satellite link to a 2Mbps terrestrial wireless link in rural Zambia.

While usage did not immediately change, performance improved soon after the upgrade. By three months post-upgrade, however, subscribers began to use the faster connection for more bandwidth-hungry applications such as video-streaming and content upload. This change in usage resulted in dramatic deterioration of network performance, whereby the average round trip time doubled, the amount of bytes associated with failed uploads increased by 222% and that of failed downloads by 91%. Thus, while an Internet access upgrade should translate to improved performance and user experience, in rural environments with limited access speed and growing demand, it can bring unexpected consequences.

Chapter 3

Technological Approach: Cellular Network Usage in Ivory Coast

3.1 Introduction

The availability of mobile networks has revolutionized the way people communicate in the developing world. Our first hand experience in rural Macha, Zambia indicates that access to cellular services is of critical importance to residents. While the reasons for adoption of cellphone technology in developing communities are not drastically different than those of the Western world, the benefits for people in these remote communities without infrastructure or other means of telecommunications is much more pronounced. Obtaining information via cell phone, as opposed to in person after travel, saves both critical time and money.

A plethora of applications that improve the wellbeing of people in remote communities, leverage cellular networks. Such applications span from health care [21,37,90,114] and education [2,117,126] to agriculture [45,115,116] and mobile banking [98]. Multiple successful projects in Africa have originated from observing user behavior in mobile or social networks. As a result of Facebook traffic analysis, Johnson et al. designed a system to facilitate local content sharing within remote rural communities [77]. Mbiti

et al. describe a system called mPesa [98] that enables transfer of money in the form of airtime in rural Kenya. The design of this system was inspired by analysis of mobile network usage in Kenya, which indicates that people tend to transfer airtime between one another as a means for payment or financial support. Follow up studies on the adoption of mPesa in Kenya show that theft decreased, as users no longer needed to carry cash.

Such projects are of critical importance to introducing new services and enhancing the wellbeing of people in under-serviced areas. At the same time, special attention should be paid in the design process of these systems to make sure that they meet an actual need in the community. Analysis of large scale datasets generated by the targeted communities naturally facilitates the identification of actual community needs.

We approach a cellular network dataset from Ivory Coast with this end in mind. The dataset provides information on an hourly basis for pairwise antenna communication over a period of five months. The dataset also features information for the personal network of 5,000 randomly selected individuals; these personal networks are called *egocentric social graphs*. Our hope is to identify unique usage patterns based on population density. We correlate cellphone usage patterns with population density and focus on aspects such as inter-antenna distance and call duration to reason about connection strength and locality of interest. Further, we analyze the egocentric social graphs hoping to identify community persistence in an attempt to motivate feasibility of information relays in user-centered cellular communication.

Previous work on geography of mobile communications focuses on traces from European countries [26, 34, 112]. In contrast, we analyze cellular network activity from a predominantly rural sub-Saharan country, where communication patterns could be different than those in Western countries due to individuals economic power, rate of adoption of cellular services and population sparsity. Other work that employs rural

and urban classification of cellular subscriptions focuses on extraction of behavior patterns of individuals living in cities vs. those in rural areas [52]. Instead, we study urban and rural area mobile usage to identify differences in call duration, distance and temporal patterns. Social network analyses using mobile traces focus on implications of network diversity [53], extracting relations [54] and community formation [112]. These works, however, are not concerned with temporal aspects of individuals' communication networks. This paper makes several contributions:

- we evaluate temporal aspects of cellular communication in Ivory Coast and identify differences with respect to mobile network utilization in the Western world;
- we identify stark differences in cellphone usage patterns in rural and urban areas;
- we find that the communication patterns between antennas in similar population densities are largely different than between antennas in different population densities;
- we design a model based on *persistence graphs* to study temporal persistence of social groups in egocentric graphs; and
- we discover that while there is a weak community persistence in egocentric graphs, there are individuals in an egocentric network that are highly persistent over time.

We start by describing our methodology of correlating population density with cellular network activity. We then describe our model for evaluation of community persistence in egocentric social graphs. We continue with extensive network analysis in section 8.5. In section 8.6 we talk about related work. Based on findings from our analysis we provide discussion of future directions in section 4.5.

3.2 Methodology

We explore the differences in communication patterns between three categories of antennas: Urban, Suburban and Rural. Apart from these typological communication patterns, we also analyze single user communication patterns over time. In particular, we look at community persistence in egocentric social graphs, whereby a subscriber of interest is centered in a graph and the periphery nodes of this graph are other subscribers with whom the central node communicates.

In this section we start by describing the datasets as provided by the telecom. We then give details about our antenna categorization. Finally, we talk about our model for egocentric social graph analysis.

3.2.1 Datasets

Our analysis is based on cellular network traces provided by one of the major mobile operators in Ivory Coast. The datasets were collected over the course of 150 days between December 1, 2011 and April 28, 2012. To assure homogeneity, the data features records only for users who were subscribed with the network for the entire capture period. Incoming and outgoing calls associated with the same session have been combined and counted as a single call. We complement this information with a dataset from AfriPop that provides high resolution population density information for Ivory Coast. A detailed description of each of these datasets follows.

Antenna-to-antenna. This dataset provides information, aggregated on an hourly basis, for the number of calls and call duration between every pair of communicating antennas in the network. Each antenna has assigned location information in the form of latitude and longitude.

Ego dataset. In this dataset the entire capture period is divided into ten equal subperiods. The dataset contains the personal communication networks of 5,000 randomly selected subscribers (egos); one network per period per ego. These personal communication networks include up to second degree neighbors of an individual and are called egocentric graphs. An edge between neighbors in these ego-graphs indicates that there was at least one call between these two users; no information for number of calls, call duration or direction is provided. Edges are drawn between (i) the ego and its first order neighbors, (ii) between two first order neighbors or (ii) between first and second order neighbors.

Population density. We use a dataset provided by AfriPop* that contains detailed spatial-geographic data for population density in Ivory Coast. Population density information is provided in ESRI Float format and can be extracted with various resolutions, the highest of which is 100 m^2.

3.2.2 Antenna classification

We categorize the antennas in the *Antenna-to-Antenna* dataset in three categories based on population density: Urban, Suburban and Rural. To do this, we employ the new European Union typology of "predominantly Rural", "Intermediate", and "predominantly Urban" areas. This typology is a modification of the Organisation for Economic Co-operation and Development (OECD) methodology that seeks to minimize distortions caused by large variations in the area of local administrative units [3]. Using the new OECD method, rural local administrative units are defined as areas with a population density below 150 inhabitants per km^2, urban are above 300 people per km^2 and suburban (or intermediate) are with population density between 150 and 300 people per km^2.

*http://www.afripop.org/

We utilize the population density information contained in the AfriPop data set and use Quantum GIS[†] to project this data as a raster layer. The AfriPop data includes population density information formatted as the number of people per 100 square meters. Since 100 m^2 is too high of a resolution with respect to typical cellular antenna coverage, we re-sample this density data at a lower resolution creating a grid of 2 km squares where the population density assigned to each square is the mean density value of the AfriPop data bounded by this square. Each square is then assigned one of the population density categories using the OECD typology. Our grid assigns population density at a resolution suitable for associating antennas with the underlying population statistics.

The majority of land area in Ivory Coast – 96.98% – is classified as Rural; 1.22% is Urban and 1.79% is Suburban. At the same time, according to the World Bank [8], urban areas account for 48.8% of the total population in the country, while the remaining 51.2% is classified as rural. Thus, almost half of the population in the country is concentrated in the few major cities, constituting a small geographic area.

3.2.3 Egocentric graphs analysis

We examine the egocentric social graphs dataset to determine persistence of social groups for each ego over time. We also analyze the likelihood that one or few nodes (users with which an ego communicates) persist over time in an egocentric graph. We hope to see persistence in both communities as well as individual subscribers. We hypothesize that such continuously-present entities can be used as information relays to strengthen information distribution amongst community members. Our analysis indicates that while community persistence is relatively low, persistent nodes indeed exist.

[†]http://www.qgis.org/

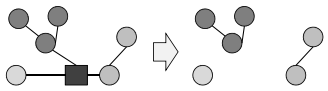


Figure 3.1: The effect of removing the ego (depicted with a square) from the egocentric social graph.

In order to extract the separate social groups of an ego, we remove the ego node from each egocentric social graph (Figure 3.1) and analyze the connected components that remain. Each connected component corresponds to one social group. Note that in the text we use the terms *connected component* and *social group* interchangeably.

After extracting the connected components we evaluate the persistence of these components over time. A connected component is 100% persistent over two consecutive periods if the nodes in this connected component are identical in the two periods. For this evaluation we define a *persistence graph* $G = (N, E, W)$ with N nodes, E edges and W weights assigned to each edge. Each node in G is a connected component labeled with the period to which it belongs. An edge exists between two connected components if they overlap in consecutive periods. The weight assigned to each edge is the *Jaccard similarity*, J, between the connected components [139]. For two sets A and B, the Jaccard similarity J can be calculated as follows:

$$J = \frac{A \cap B}{A \cup B} \tag{3.1}$$

The Jaccard similarity ranges between 0 and 1, where 0 indicates no overlap and 1 indicates full overlap.

Figure 3.2 presents an example of building the persistence graph for a single ego over three consecutive periods. The left-hand side of the picture presents the set of neighbors in each of the three periods. The social groups comprised by these neighbors are color-coded. The right-hand side of the picture presents the resulting persistence graph. Each node corresponds to a connected component (CC) in a given period. In

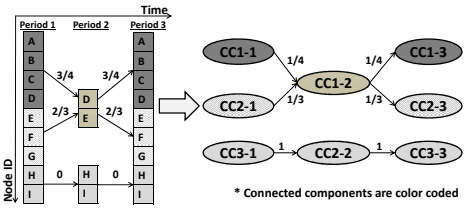


Figure 3.2: Building a persistence graph.

the figure nodes' labels are of the format CCID-PeriodID. Edges exist only between connected components that overlap fully or partially in consecutive periods. There is no edge between connected components that persist over non-consecutive periods (e.g. there is no edge between node "CC1-1" and node "CC1-3").

Our persistence analysis is based on the described persistence graphs and consists of two parts. First, we analyze the in- and out-degree distribution of the nodes in the persistence graph. We note that if the social groups of an ego persist over time, all the nodes in the persistence graph should have in- and out-degrees of either 0 if the node belongs to the first or last period, or 1 if the node is in the intermediate periods. In cases where social groups do not persist, nodes can have a degree of 0 if the corresponding social group does not re-appear in following periods. Nodes can also have in- and out-degrees larger than 1 if social groups merge or split in consecutive periods.

We attempt to quantify the level to which social groups overlap by considering the weights of the edges in the persistence graphs. As detailed earlier, edges are drawn between nodes that overlap fully or partially in consecutive time periods. The weights assigned to these edges are the Jaccard similarity between the nodes connected by these edges. For each transition between period t and period $t+1$ we find the normalized Jaccard similarity $\hat{JS}^{(t,t+1)}$ between these periods: that is the sum of edge weights

$W_i^{(t,t+1)}$ divided by the number of edges $|E^{(t,t+1)}|$ between the two periods.

$$\hat{JS}^{(t,t+1)} = \frac{\sum_{i=1}^{|E^{(t,t+1)}|} W_i^{(t,t+1)}}{|E^{(t,t+1)}|} \tag{3.2}$$

We then find the average Jaccard similarity for the entire persistence graph by summing the normalized Jaccard similarities and dividing this sum by the number of period transitions K.

$$\bar{JS} = \frac{\sum_{j=1}^{K} \hat{JS}_j^{(t,t+1)}}{K} \tag{3.3}$$

Informally, the higher the average Jaccard similarity, the more persistent the social graphs of an ego are over time.

We present our results for social group persistence in Section 3.3.7.

3.3 Network Analysis

We begin our analysis by mapping antennas to population density and discussing general trends in antenna utilization. We then investigate temporal trends in mobile communication in general and across areas of different population density types. We expect to observe regular temporal trends along weekly and monthly intervals, with Rural areas having distinctive patterns from those of Urban areas. We also explore trends related to population density; we expect to see differences in call duration and call frequency based on the population density of the sender's geographical area. Next, we analyze how the inter-antenna distance impacts the call duration and call frequency. Finally, we examine patterns in social groups, hoping to observe persistence in social groups over time.

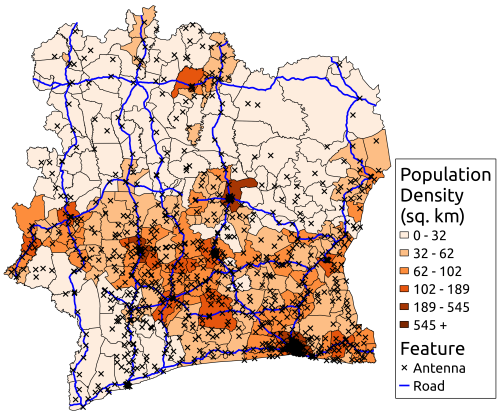


Figure 3.3: Cellular Antennas in Ivory Coast.

Table 3.1: Antenna Density Classifications

Classification	Antenna Count	Source Calls
Rural	528	146,481,488
Suburban	90	21,529,115
Urban	598	331,630,147
Unknown	15	65,393,926

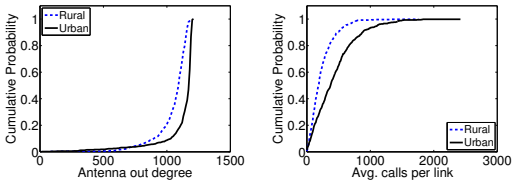
3.3.1 Antenna activity

We study patterns of mobile communication in Ivory Coast by associating antennas with their geographic location and population density. The resultant mapping of antennas to location is shown in Figure 3.3. The figure presents average population density per sub-prefecture with overlaid antennas. It is evident that antennas are densely clustered in urban locations while more sparsely located in predominantly rural regions. We also find that high activity antennas are often located along major transportation corridors.

We explore the relationship between the population density of a sending antenna and the average number of outbound calls associated with the antenna. Because of the predominant use of "Calling Party Pays" (CPP) policy in sub-Saharan Africa, we focus on the number of outbound calls rather than incoming calls [49, 102]. Due to the

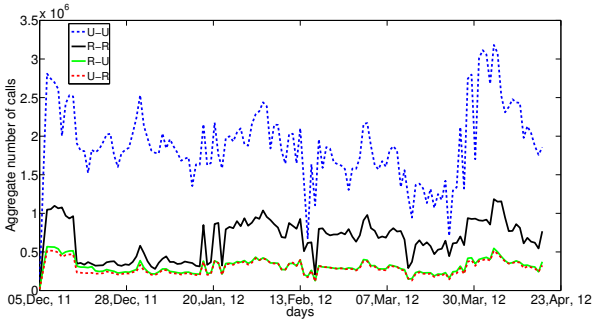
CPP policy, we anticipate a larger mean number of outbound calls from antennas in high population density areas, which coincide with Ivory Coast's financial district and center of commerce. In Table 3.1 we show the number of antennas that fall into each of the classifications as well as the total number of calls originated from each antenna type. As expected, while the number of Rural and Urban antennas is almost the same, the amount of calls originated by Urban antennas is more than twice as large as those originated by Rural antennas. Very few antennas and calls are classified as Suburban. This is consistent with the fact that the subset of Suburban population is very small in comparison with Urban and Rural. In the rest of this paper we will focus our evaluation on activity associated with Rural and Urban antennas.

We finalize our antenna activity evaluation by examining the outbound communication trends per antenna pair in Urban and Rural areas. Again, we focus on call originators. In particular, we evaluate over the entire period, (i) the outbound degree of each antenna, meaning the number of connections each antenna establishes with other antennas, and (ii) the sum outbound weight of each antenna expressed as the sum of the number of calls originated on each outbound link. We present our results in Figure 3.4. Figure 3.4(a) plots the CDF of out degree of antennas in Urban and Rural areas. As we can see, the mean outbound degree for Rural antennas is lower than that for Urban. This indicates that Rural antennas tend to communicate with fewer antennas than Urban. We then evaluate the average strength of antenna to antenna links by examining the average weight of the outgoing links associated with a given antenna; link weights are assigned according to the number of originated calls. Figure 3.4(b) plots a CDF of average link weight for Rural and Urban antennas. The average weight for Rural is twice as small as for Urban, which implies that on average more calls originate from Urban than from Rural antennas.

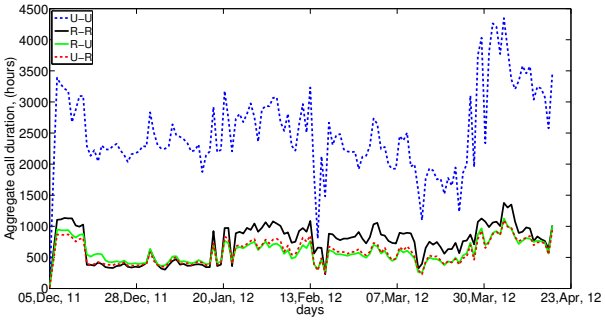


(a) Antenna out degree (b) Antenna out weight

Figure 3.4: Antenna activity.



(a) Aggregate number of calls



(b) Aggregate call duration

Figure 3.5: Mobile network activity over time.

3.3.2 Temporal usage patterns

We evaluate the cellular network activity patterns over the entire capture period. We categorize calls by originating and receiving antenna type based on the antenna categorization from section 3.2.2. In Figure 3.5 we plot aggregate number of calls and call duration per day for four categories of antenna pairs. In the legend, U and R

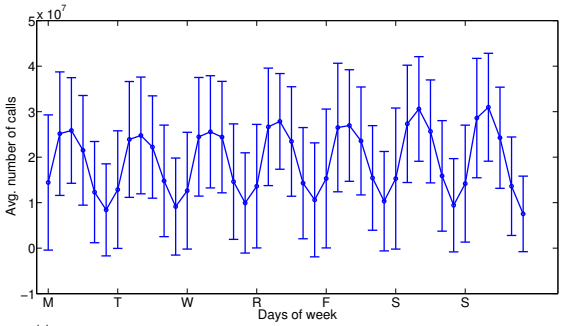


Figure 3.6: Weekly patterns.

stand for Urban and Rural, respectively. As the figure shows, there is no distinctive call pattern on a weekly or monthly basis; instead, subscriber activity seems to be widely correlated with events in the country. We hypothesize that the peak near the beginning of the period coincide with the weeks before and after the parliamentary elections on December 11th, 2011, while the second peak is most likely traffic related to New Years. The increased utilization from the end of March through April is likely associated with the military coup in Mali and the ECOWAS[‡] summit that took place in Abijan, Ivory Coast. Such irregular usage pattern is very different than what had been observed in cellular network traces from the Western world [89].

In Figure 3.5 we also illustrate that the calling patterns in Rural areas differ from these in Urban areas. As the figure shows, calling patterns for all four categories follow similar trends, where the number of calls and the aggregate call duration between Urban antennas is about three times higher than between Rural antennas. We note that while the number of Rural to Rural calls is larger than the number of Rural to Urban and Urban to Rural, the aggregate call duration for these three categories is the same (Figure 3.5(b)). This result indicates that while calls between Rural residents occur more often, they are shorter in comparison to calls between Urban and Rural residents.

[‡]`http://www.ecowas.int/`

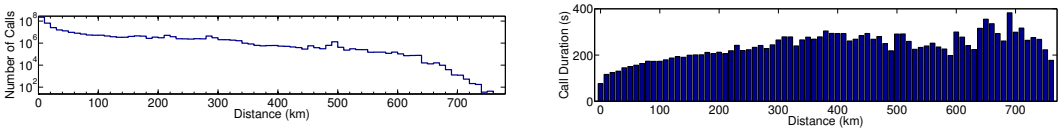
Next we study weekly communication patterns. To extract weekly behavior, we average the number of calls and call duration over the entire capture period for each week day. Figure 3.6 presents our results. Each point on the plot presents an average over four hours over all occurrences of each day of the week (that is the first data point from the graph presents the average number of calls for the hours from Midnight to 4 AM for all Mondays in the capture period). The figure clearly presents diurnal pattern of network activity with slight increase over the weekend. However, the standard deviation of this graph is very high, indicating that the network activity varies dramatically over the observed period.

3.3.3 Antenna distance

We investigate the relationship between call distance and the average duration of calls. We calculate the distance in kilometers between all pairs of antennas with known geographic location using the Haversine formula [129] with mean Earth radius of 6,372.80 km. We group connection distances into the nearest 10 km in order to calculate aggregate statistics for each group.

First we examine trends in frequency of communication as a function of distance between calling parties, shown in Figure 3.7(a). The figure plots number of calls in log-scale on the y-axis and distance on the x-axis. We see a long tail distribution of number of calls over distance, whereby antennas in close proximity tend to have many more calls between one another in comparison with antennas that are further apart.

Next we evaluate the impact of distance between antennas on the mean call duration. We find the mean call duration by grouping all antennas within a certain distance from one another and dividing the total call duration by the number of calls. The impact of distance between source and destination antenna on mean call duration is shown in Figure 3.7(b). While the number of calls in close proximity is much higher, the



(a) Number of calls vs. antenna distance (b) Mean call duration vs. antenna distance

Figure 3.7: Communication patterns as a function of the distance between antennas.

average call duration is lower when communicating antennas are nearby, and increases as connection distance grows. We hypothesize that the reason for such increase in call duration is that calling parties who are further apart have fewer opportunities for in-person interactions, thus they tend to talk longer over the phone. Lastly, note that with relatively few call records for distances greater than 500 km, more noise is introduced into the graph.

3.3.4 Call typology classifications

We investigate the potential correlation between population density and calling patterns by associating antennas with the corresponding local population density. This process yields antennas denoted as Rural, Suburban, Urban, or Unknown (for the antennas which have no geographic location). We process the *Antenna-to-Antenna* set to classify call records by each typology source and destination pair in order to investigate potential communication patterns. In this analysis we do not consider records for antennas with no geographic data or records without valid antenna IDs.

We start by analyzing the distribution of antenna pairs in four categories: U-U, R-R, U-R and Other. Note that in this classification we do not consider directionality. The Other category contains antenna pairs featuring Suburban antennas as well as such that are unclassified. As seen in Figure 3.8, the majority of connections are classified as U-U connections. This is followed by 20% of connections classified as R-R. Mixed links of

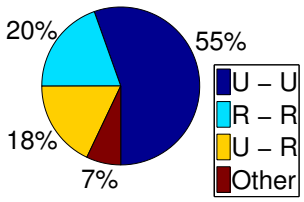
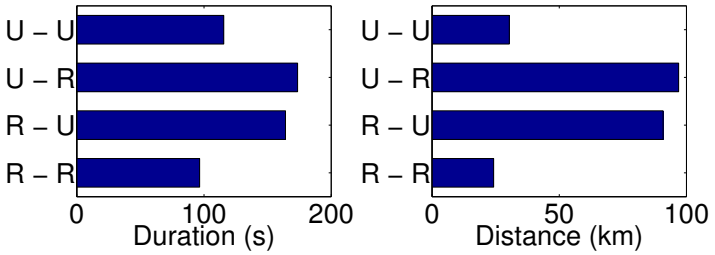


Figure 3.8: Classification of communication between antenna pairs.

R-U account for 18% of the total. The relatively small fraction of calls between mixed antennas can be explained with the high locality of calls observed in Figure 3.7(a), according to which the majority of calls occurs between antennas which are a few tens of kilometers apart.

Next, we search for differences in mean call duration across the connection classifications and show results in Figure 3.9(a). We find that the two call classifications with the longest mean call duration are Urban to Rural and Rural to Urban. An observable phenomenon is that calls confined to the same source and destination density type are noticeably shorter on average compared to calls between mixed types. Given our prior finding of the relationship between call distance and average duration, we posit that the majority of calls that do not cross classification boundaries are confined to a smaller geographic region. For instance, we believe Urban to Urban calls are more likely to be sourced from and destined for the same urban area. Lastly, an interesting observation is that calls originating from Urban antennas generally have a longer duration for any destination type. This is likely due to the common policy of "Calling Party Pays" and higher buying power of individuals who reside in urban areas.

This trend leads us to look at the average distance between connecting antennas associated with each connection density classification type, shown in Figure 3.9(b). The longest average distance between connecting antennas occurs in the Rural to Urban and Urban to Rural cases. The shortest average distance occurs between similar source/destination connections. In the case of Urban to Urban this is likely due to



(a) Duration (b) Distance.

Figure 3.9: Mean call duration and mean call distance for connections of different types.

the fact that there are only two Urban areas, which are possibly closer to one another in comparison with an arbitrary Urban to Rural case. The case for Rural to Rural, however, is more interesting. While the vast majority of the country is Rural, thus statistically Rural to Rural communications can cover long distances, the Rural to Rural patterns presented in Figure 3.9(b) indicate that people in Rural areas who call one another tend to be in close proximity. This indicates high locality of interest in Rural to Rural cellular communications in Ivory Coast.

To further explore this locality of interest, we investigate associating call patterns and population density for calls that have the same source and destination antenna ID. We find that in 57% of all Rural to Rural calls are sourced from and destined to the same antenna. We posit that this is due to fewer available antennas in predominantly rural areas. Furthermore, the coverage area of a single antenna in rural settings is typically larger (up to 35 km), which means that a higher proportion of local users are associated with the same antenna. Nevertheless, this high percentage of same antenna calls confirms that cellular communications in Rural areas are very local. Interestingly, Urban connections sourced from and destined to the same antenna represent 23% of all Urban to Urban calls. We believe that the higher density and smaller cell range of Urban antennas provides more diverse antenna association possibilities for users.

3.3.5 Transportation infrastructure

A close analysis of Figure 3.3 shows that a large fraction of antennas in areas of low population density are situated in close proximity to major transportation corridors in Ivory Coast. In our current antenna typology, these antennas close to roads are categorized as Rural antennas. However, we expect that the usage patterns of antennas associated with transportation corridors will differ significantly than those located in Rural residential areas. Thus, we divide the Rural category in two subcategories: Transportation and Rural-residential, where an antenna is labeled Transportation if it is within 5 km of a highway. As a result we find that 51.7% of the antennas the were originally classified as Rural are associated with road infrastructure. Of note is that a Transportation antenna can be used by both travelers as well as Rural residents, which is why all results except for the ones presented in this section, feature both Transportation and Rural-residential antennas in the Rural antenna type.

Based on our new classification, we evaluate communication patterns in terms of three indicators: (i) call duration, (ii) number of calls and (iii) percentage of calls where both the originator and the terminator are associated with the same antenna. We find that the average call duration for Rural-residential to Rural-residential calls slightly decreased from 95.5 seconds to 84.8 seconds. More drastic, however, was the change in the same antenna calls and the average distance between calling parties. While previously the same antenna Rural calls were 57%, in the case of Rural-residential communication the same antenna calls increases to 70%. Accordingly, the average distance between Rural to Rural calls decreased from 24.2 km to 13.4 km after the Transportation antennas were removed from the Rural typology. This result once again confirms the strong locality of interest in cellular communications in rural Ivory Coast.

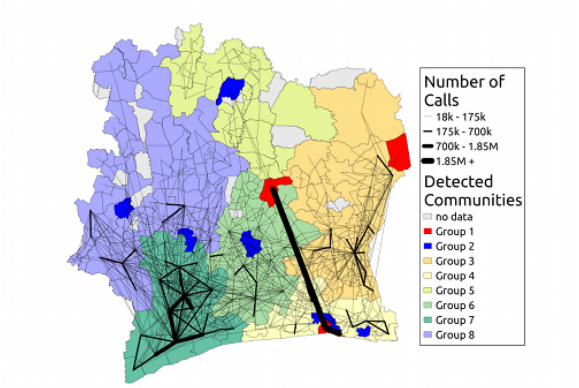
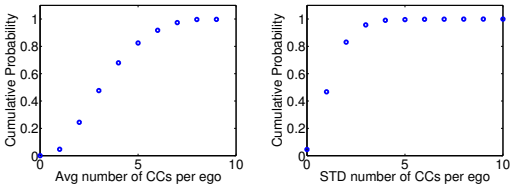


Figure 3.10: Weighted connectivity map with detected communities.

3.3.6 Community discovery

We examine the antenna connectivity data as a weighted graph in order to seek out underlying community structures using the modularity algorithm proposed in [27]. Communities are defined as subsets of highly connected nodes in a graph [108]. Nodes in our graph are represented by sub-prefectures with edges indicating connectivity between antennas lying within the two sub-prefectures. Edge weights are represented by the total number of calls between a given sub-prefecture pair. Figure 3.10 shows detected communities using a resolution value of 1.15 for the modularity algorithm. The resolution value is data-dependent and determines the number of detected communities for a given data set. We find 1.15 to be an optimal value for our data set in order to avoid too many or too few communities. The figure edges are weighted based on the number of calls with only the top 2% of edges displayed for readability. Sub-prefectures are shaded to indicate community membership. Sub-prefectures labelled as “no data” are those in which an antenna does not exist, therefore they are not considered a part of the graph.

After applying the community detection algorithm we find that with a few exceptions, communities are largely based on close proximity within geographic area. As the

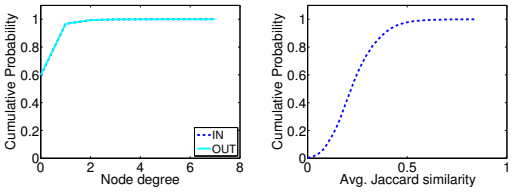


(a) (b)

Figure 3.11: (a) The number of connected components (CCs) per ego and (b) the standard deviation of the number of connected components per ego over the observed period.

figure shows, "Group 1" includes two of the largest cities (Abidjan and Bouaké) despite the fact that they are geographically non-contiguous. We believe the easternmost sub-prefecture included in "Group 1" is an anomaly given that only two antennas exist in the region. Perhaps the sub-prefecture includes a facility in which workers often contact major cities. Most interestingly, "Group 2" includes multiple geographically dispersed medium-sized cities as well as the area surrounding Abidjan. "Groups 1" and "2" indicate that as population density increases, geographic distance becomes less of a factor in relative "connectedness" in the network. The remaining groups largely consist of rural regions that are geographically similar.

The results support our observation of a distinct divide between rural and urban regions. Rural communication patterns are largely regional and tightly connected with the surrounding area. On the other hand, urban areas are tightly connected with other urban areas and not necessarily with nearby rural areas. As predominantly disconnected rural areas are often the target of connectivity research, this insight can be used to inform the design of any solution. While global connectivity is the ultimate goal, local and regional telecommunications connectivity can clearly benefit users in a localized rural area.



(a) (b)

Figure 3.12: (a) The in- and out-degree of nodes in all persistence graphs and (b) the average Jaccard similarity for each persistence graph.

3.3.7 Egocentric graphs

Now that we understand regional connectivity patterns, we take a deeper look at egocentric social graphs provided in the *Ego* dataset. Our analysis focuses on persistence of social groups with which individual egos communicate. We regard this analysis as preliminary work on identifying persistent neighbors within one's social network who can serve as reliable information relays.

First we provide high level analysis of the average number of social groups with which each ego communicates over the entire capture period from December 2011 to April 2012. For this analysis we sum the number of connected components that appear in each two-week period and divide this sum by the number of capture periods. Figure 3.11(a) plots a CDF of the average number of connected components for each ego. While the average number of components across egos spans from 1 to 10, the majority of egos – 68%, have between 2 and 5 connected components on average. Further, we examine how the number of connected components deviates for each ego. Figure 3.11(b) plots a CDF of the standard deviation of the number of connected components per ego over the observed period. Almost half of the egos (47%) have standard deviation of less than 1, while 96% of all the egos have standard deviation of less than 4. This indicates that the number of connected components in an egocentric graph remains relatively constant over time.

Next we analyze the persistence of these social groups over time. First, we look at the in- and out-degree distribution of nodes in the persistence graphs. As detailed in Section 3.2.3, a node in period t has in- or out-degree of 0 if it belongs to the first or last observed period or if it does not overlap with any node from the preceding $(t-1)$ or the following $(t+1)$ period. Nodes have in- and out-degree of exactly 1 if they persist over time, and degree larger than 1 if they split or merge over consecutive periods.

We calculate that out of all the nodes in all persistence graphs, 9.49% belong to the first period (i.e. have in-degree of 0) and 8.93% belong to the last period (i.e. have out-degree of 0). At the same time Figure 3.12(a) indicates that in nearly 60% of the cases nodes have in- or out-degree of 0. This means that about 50% of all the social groups that we observe, and which were not associated with the first or last period, did not occur in the preceding and following periods. 40% of the nodes have in- or out-degree of 1, indicating that 40% of the social groups persisted in consecutive periods. Only about 3% of the cases have in- or out-degree larger than 1; social groups rarely split or merge over consecutive periods.

This result indicates an important quality of the observed egocentric social graphs: there are two distinctive types of social groups with which an ego communicates – (i) those that likely occur only once (in- and out-degree is 0), and (ii) those that likely persist over time and strictly correspond to one social group from the preceding and one social group from the following period. The former group can be associated with one-time calls, for example calling to schedule a doctors appointment, while the latter can be associated with calls recurring over time, such as these between relatives and friends who stay in touch.

We continue our evaluation of social group persistence by analyzing the weight of edges (representing the similarity) of social groups in consecutive periods. We leverage the average Jaccard similarity metric as defined in Section 3.2.3; the closer this similar-

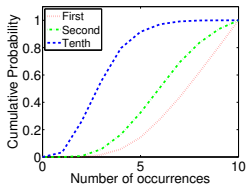


Figure 3.13: Number of occurrences of the first, second and tenth most frequent neighbor.

ity is to 1, the larger the overlap between social groups in consecutive periods. Figure 3.12(b) plots a CDF of the average Jaccard similarity for the 5,000 ego-centric graphs. The median of this CDF is only 0.22, which means that on average the overlap of social groups over time is relatively small – about 22%.

Finally, we evaluate the frequency of occurrence of the neighbor that appears most often in the social network of an ego. For this evaluation we count in how many of the ten observed periods each neighbor appears. We then sort the neighbors in decreasing order of appearance frequency. We compare the first, second and tenth most frequent neighbors to determine whether there are groups of neighbors that appear more often and what is a typical size of such groups.

Figure 3.13 presents our results. The median value for the first top neighbor is 8, while for the second and the tenth top neighbor it decreases to 6 and 3, respectively. This means that in 50% of the cases, the most frequently occurring neighbor exists in 8 out of the 10 observed periods. These results indicate high persistence of at least one neighbor in the social graph. At the same time, a group of two most persistent neighbors would appear ten times in only 6.8% of the cases, which indicates that a group of most persistent neighbors would typically have very few members.

3.4 Related work

Analysis of mobile network traces provides a unique opportunity for large-scale verification of socio-economical models that were previously derived and studied on much smaller scales. The correlation of mobile network traces with other data provides for realistic analysis of problems spanning from epidemiology [147] to individuals' economic prosperity [53], relational dynamics [54] and gender divides [28]. Previous work focuses on epidemiology informed by human mobility patterns in cellular networks. For example in malaria tracking, mobile network traces with subscriber location information are correlated with malaria prevalence information. This correlation enables quantification of the impact of human mobility on the spread of malaria: see [147] and references therein.

Previous research related to our work can be divided in three categories: (i) geography of mobile communications, (ii) mobile communication patterns informed by population density and (iii) social interactions analysis.

Geography of mobile communications. Previous work explores geographical aspects of human social interactions in mobile networks. Onnela et al. analyze one month of mobile traces from a European country to discover relations between community formation and geographical distance between community members [112]. Other work based on European traces [34] analyzes patterns in pairwise communication to determine whether call duration and frequency change based on the physical distances and opportunities for in-person communication between individuals. Blondel et al. apply a grouping algorithm on a cellular network trace from Belgium to extract the geography of mobile communication in the country [26]. The presented results indicate that grouping of mobile call interactions is both region and language dependent. In contrast our work focuses on mobile traces from sub-Saharan Africa, where communication

patterns could be different than those in Western countries due to specifics related to population sparsity, adoption of cellular services and individuals purchasing power.

Population density. Previous work on mobile network analysis informed by population density focuses on studying behavioral patterns in rural and urban areas [33, 52, 59]. Eagle et al. employ mobile data to study differences in human behavior related to personal network topologies and travel patterns of individuals living in cities as opposed to those living in rural areas [52]. In contrast, we study urban and rural mobile usage to identify differences in call distance, duration and temporal usage. Our findings make a strong case for locality of interest in rural cellular communications in Ivory Coast.

Social analysis. Social network analyses based on cellular traces focus on implications of network diversity [53], extracting relations [54] and community formation [112]. Studies demonstrate that diversity of one's mobile social network influences socio-economical prosperity [53]. Other work extracts biases in self-reported friendships by comparing characteristics of self-reported relationships with those extracted from cellular traces [54]. These studies, however, are not concerned with variability of social networks over time. In contrast our analysis explores temporal trends of cellular communication in individual subscribers' communication networks and provides insights on community persistence in egocentric social graphs.

Dynamic graph mining. In the area of dynamic graph mining, research has focused on evolutionary community detection [84], conserved relational states [16] and high-scoring dynamic subgraphs [29]. Bogdanov et al. propose a method to identify the highest-scoring temporal subgraph (e.g. most congested road segment) in a dynamic network [29]. Our analysis is different, as we seek to summarize the persistence in different egocentric networks without observing the whole graph at a time. Other work mines relational patterns in a dynamic network [16] in order to detect maximal evo-

lution paths in time-evolving networks. While this work utilizes a model for tracking similarity that is similar to our persistence graphs, the proposed scheme is only concerned with full overlap of graph entities over time. In contrast, our method captures partial overlaps and allows for fine-grained analysis of community persistence.

3.5 Discussion and conclusion

Our analysis of cellphone traces from Ivory Coast indicates that temporal usage patterns differ than those in European countries [89]: while studies in Europe show persistent diurnal and weekly patterns over time, such patterns are not typical to Ivory Coast. The reason for such difference could be related to several factors. First of all, while in Europe access to cellular communication is a regular part of daily life, cellphone services in developing areas are still being deployed at large and the number of subscriptions is actively growing. In this sense the cellphone user population and usage habits are still being established, which could lead to highly-dynamic utilization trends. Secondly, because urban areas, which are the center of governance and commerce, are the main contributor to cellphone activity in Ivory Coast, the cellular network utilization is likely largely dependent on socio-political or cultural events in the country.

Urban areas are the main contributor to cellular activity in Ivory Coast: 55% of all calls occur within urban areas and another 9% of the activity is initiated by urban areas. We speculate that this could be due to several reasons. First, people in urban areas have more coverage than those in rural areas: while only 1.22% of the territory of Ivory Coast is urban, 48.8% of the population and half of the cellular antennas reside in urban areas. Furthermore, it is possible that people in urban areas have higher buying power than those in rural areas. In line with this assumption and the CPP policy in sub-Saharan Africa, we find that Urban-to-Urban calls and mixed calls initiated by urban

residents are longer on average than those originated by rural residents. Of note is that the paying ability is not the only factor that influences cellular usage; factors such as adoption, needs and communication habits can also influence usage patterns. Thus, further analysis that involves fine-grained per-subscriber call information and air time purchasing information would help clarify calling trends.

We show that urban and rural usage patterns in Ivory Coast differ significantly. For instance, while the total number of rural and urban antennas is comparable, rural antennas tend to connect with fewer antennas than do urban antennas. Furthermore, a rural antenna originates fewer calls on average than an urban antenna. At the same time the locality of interest in rural communication is much higher than in urban communications: 70% of the Rural-residential calls happen in the vicinity of the same antenna. This means that people in rural residential areas tend to call other rural residents in close physical proximity. While global connectivity is the ultimate goal, this finding makes a strong case for the feasibility of local cellular communication solutions such as the ones discussed in [20, 64, 153]. The actual applicability of such solutions, however, depends not only on availability of low-cost power-saving technology but also on licensing. The current licensing schemes are extremely conservative, costly and oblivious to return of investment. We argue that these licensing aspects are the main factor preventing entrepreneurs from deployment in rural remote areas.

Lastly, we evaluate the persistence of social groups in egocentric social graphs. With this analysis we provide preliminary insights to inform future work on extraction of information relays from egocentric cellular social networks. Such information relays can be used to strengthen information channels between educators and students or health workers and patients. Our first-hand experience in rural health care indicates that improvement of information channels is indeed needed since health care services such as immunizations are often available; however, it is difficult to bring information about

availability to the interested patients. Analyzing 5,000 random users from Ivory Coast, we find that on average an egocentric network has four social groups; this number is stable over time. We also find that 50% of the observed social groups did not occur in the corresponding preceding and following periods. At the same time, less than 1% of the communities split or merge over time; thus communities that do persist tend to be independent of one another. This means that multiple information relays can be selected from each independent community, to increase the chance that information will flow to the ego. Finally, we observe that on average there is a 22% overlap of social groups over time. Persistence of a subscriber in one's social group likely means that there is a stronger personal connection between the ego and the corresponding subscriber. Thus, such subscriber persistence needs to be taken into account in the process of relay selection. In order to devise models for information relay extraction, however, more detailed information is needed that contains not only location but also frequency and duration of interaction. Availability of such information will enable true extraction of individuals that are strongly connected to an ego and can serve as reliable information relays.

3.6 Acknowledgements

The material in this chapter is based on joint work with Paul Schmitt, Morgan Vigil and Elizabeth Belding.

Chapter 4

Technological Approach: Internet Performance and Usage in Rural Zambia

4.1 Introduction

Access to the Internet is critical for improving the wealth of nations and promoting freedom. Bright examples of advancements facilitated by Internet access span from democratic changes [19], to government [107], e-learning [132] and health care [55]. Broadband Internet access, however, is still largely unavailable in developing countries with only 6% of the population having broadband connectivity [72], the majority of which is in urban areas.

Recent efforts to bring connectivity to rural areas of the developing world utilize asymmetric satellite or other low-bandwidth wireless links [96, 137]. At the same time the bandwidth demand of online applications is increasing; for example the average web-page size has grown 110 times since 1995 [85]. As a result, residents of developing rural regions access the web with inadequate conectivity for the bandwidth requirements of modern content. These opposing trends in content growth and limited

capacity render Internet access frustrating or even unusable [50,78] in many developing areas.

Previous work on traffic analysis shows a "strong feedback loop between network performance and user behavior" [78], whereby residents in bandwidth-constrained environments tend to focus more on bandwidth-light applications such as web-browsing, as opposed to those in bandwidth-rich environments which enable multimedia streaming, content upload, and real-time user interaction. In the face of limited bandwidth, the failure rate of uploads is high [79], discouraging rural residents from contributing to the Internet content and resulting in consumption of largely Western content [143]. Thus, while recognizing the potential benefits of the Internet, residents of developing regions express concerns that the flood of Western culture, coupled with decreased ability to document and transfer their own traditions, threatens the existence of local cultures [142].

The focus of our work is in Africa, where the increased fiber-optic capacity [1], coupled with higher-bandwidth, lower-latency technologies such as terrestrial microwave wireless gives hope for improved Internet access in remote areas. In this paper we study the implications of an Internet access upgrade from a satellite to a microwave terrestrial link on the performance and Internet usage in the rural community of Macha, Zambia. To the best of our knowledge, this is the first real-world comparative study of pre- and post-upgrade Internet usage and performance. As such, our dataset offers a unique opportunity to study the change in user behavior and Internet usage following an eight-fold increase in access bandwidth. We evaluate a total of three months of usage: one month before the upgrade, one month after the upgrade and one month three months later. Our results show that while usage did not change immediately, application performance improved. However, as time progressed subscribers began to change their Internet usage behavior, which ultimately resulted in network performance degradation

and subsequent deterioration of user experience. The Internet access upgrade broadened users' abilities to access content, use online applications, and express themselves on the Internet. At the same time our results make a strong case that one should not assume that advanced technologies and higher access speed grant better experience and increased adoption of the Internet in rural communities; rather one should carefully consider the evolution of usage and performance in order to assess the actual impact and adoption of Internet technologies.

4.2 Network Analysis

Table 4.1: General TCP statistics averaged over each time period.

	Total GB	Packets (x10^6)	Control packets (%)	Avg. Window (kB)	Avg. RTT (s)	Retrans-missions (%)
Pre	123	373	56.59	38	0.1436	1.12
Post	163	338	47.69	52	0.1085	1.09
LT	210	432	49.72	62	0.3190	1.16

We evaluate the network performance and usage for three months. We select one month immediately before (which we call Pre-upgrade) and one month immediately after the upgrade (Post-upgrade) to measure the short term impact on the network usage and performance. We also evaluate one month of traffic approximately three months Post-upgrade to determine whether performance changed as time progressed. We call this time period Long-term.

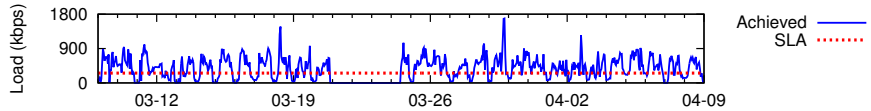
We start by describing our traffic collection methodology as well as our approach to calculating evaluation metrics. We then continue with detailed results from our network analysis. We first focus on overall network performance analysis, which indicates that the majority of traffic traversing the network is TCP (93%). We, thus, focus our analysis on TCP performance following the increased bandwidth. We describe trends in uplink and downlink performance of TCP flows, and we identify the most popular applications

based on TCP port number. We then assess the success and failure rate of TCP flows. We conclude our TCP analysis by outlining performance trends in Windows and Linux machines. We then switch to evaluation of network usage focusing on popular URIs. We conclude by analyzing the "worldliness" of network flows initated in Macha in an effort to determine whether Machans started using more global services once they had better Internet access.

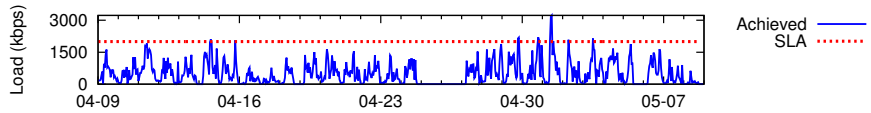
Pre-upgrade the network was typically saturated, resulting in high round trip time (RTT), congestion, and aborted sessions. Post-upgrade, we saw a decrease in the number of retransmissions and RTT due to improved network performance and movement away from the saturation point. By three months after the upgrade, the traffic had increased once again to saturation. Our analysis shows a difference in network performance and utilization Post-upgrade and Long-term: while Post-upgrade user behavior did not change, automatic programs, such as software updates, were suddenly able to complete, which resulted in an increase in traffic demand. In Long-term, subscribers utilized the faster Internet access for more bandwidth-hungry applications such as video streaming. Once the saturation point was reached in Long-term, network performance deteriorated, but was still better able to support bandwidth-intensive applications than Pre-upgrade. We describe these network usage and performance patterns in detail in the following sections.

4.2.1 Methodology

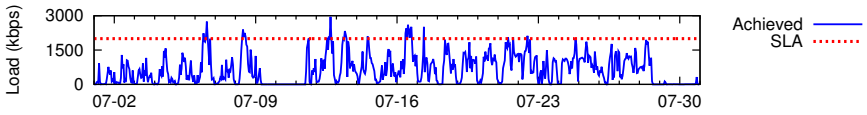
We capture traffic at the Internet gateway in Macha. As shown in Fig. **??**, we connect a monitoring server to the switch that bridges the Internet gateway and Macha's WLAN. We configure a mirror port at that switch, which allows us to capture all the traffic traversing the WLAN. With user consent, we capture packet headers and store traces on the monitoring server. During our last trip to Macha in Summer 2012, we of-



(a) Pre-upgrade



(b) Post-upgrade



(c) Long-term

Figure 4.1: Traffic load over time.

floaded the collected traces to an external hard drive and brought them to our research facility for offline analysis.

We now describe our methodology for extracting metrics from the collected network traces. In our evaluation we use metrics such as TCP window size, RTT, TTL and retransmissions. We extract these metrics by running `tshark` in an offline mode on the collected traces. For our flow analysis we developed a tool that reassembles unidirectional flows from a list of packets based on packet signature (source IP, source PORT, destination IP, destination PORT, timestamp). In the process of flow reassembly we count the number of packets and bytes associated with this flow and calculate its duration. We calculate the packet Inter-Arrival time (IAT) as the difference in time of consecutive packets. In order to obtain bidirectional flows, we then combine the unidirectional flows based on flow signature and timestamp.

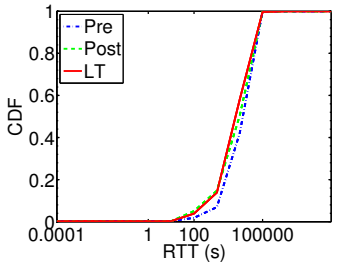


Figure 4.2: RTT.

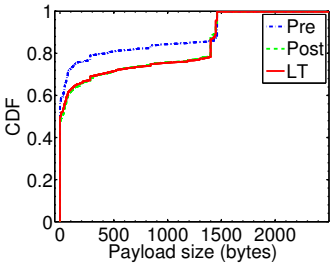


Figure 4.3: Payload size.

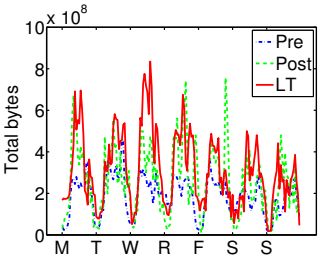


Figure 4.4: Bytes by day.

4.2.2 Overall network performance

Traffic load. We start with evaluation of the traffic load. We calculate the load as the aggregate number of bits that traverse the gateway each hour divided by the number of seconds in an hour; our results capture the average combined uplink and downlink rate. We find that the average traffic load Pre-upgrade is $367.3kbps$, Post-upgrade is $495.3kbps$ and Long-term is $648.1kbps$. Fig. 4.1 plots over time the traffic load averaged per hour in blue and the Service Level Agreement (SLA) with the Internet provider in red*. In the period before the upgrade, the demand frequently exceeded the SLA of 256kbps. This is less often the case for the period immediately after the upgrade, as users have not yet adapted to the increase in bandwidth. However, three months after the upgrade the demand often approaches the SLA. As detailed later in our analysis, this is likely due to changed usage patterns whereby users began to access more bandwidth-hungry applications once more bandwidth was available. The gaps in the plots correspond to time periods in which traffic captures were unavailable due to power or network outages.

*Note that while the guaranteed speed was 256kbps, bursts of up to 1Mbps were possible depending on link utilization. This is why the actual traffic load Pre-upgrade consistently exceeds the SLA of 256kbps.

General trends. We continue our evaluation by discussing general trends over the three observed periods. Table 4.1 presents a detailed look into performance. As we can see, the total bytes that traversed the gateway nearly doubled in the course of three months. The total number of packets dipped Post-upgrade, as the same traffic demand was first accommodated with fewer retransmissions. As time progressed usage changed, which resulted in drastic increase in the total bytes traversing the gateway and a corresponding increase in the number of packets.

A similar trend is observed in RTT. While immediately after the upgrade the average RTT decreased by about $35ms$, it nearly tripled as time progressed. We explore the RTT dynamics in more detail in Fig. 4.2, which plots a CDF of RTT for the three periods. We observe a long-tail distribution of RTT in Post-upgrade and Long-term performance; however, the median values of RTT for those two periods are lower than those observed Pre-upgrade. As we will see later in our analysis (in section 4.2.5), the long-tail distribution of RTT after the upgrade is due to changed browsing habits and tendency to use services that are physically further away (such as streaming video from servers abroad). We provide in-depth discussion of usage patterns in section 4.2.4 to validate our hypothesis.

We analyze payload size in Fig. 4.3. We see a clear bi-modality [133] of payload size, which is due to the prevalence of either control packets with 0 bytes payload or data packets with payload of about 1500 bytes. Clearly, the percentage of large data packets Post-upgrade as well as Long-term increased. We also see an increase in the average TCP window size (Table 4.1), which allows more packets to be sent in the network before an acknowledgement is received. This increased TCP window size is critical to improved TCP performance as it translates to higher achieved throughput.

We next measure the overhead, focusing on the percentage of the total packets that are due to retransmissions and control packets (e.g. TCP control packets are ACK,

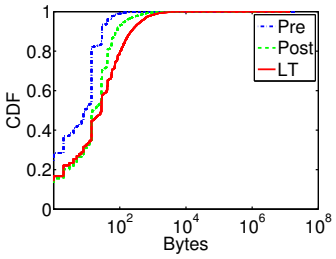


Figure 4.5: Bytes in flight.

SYN, FIN). The payload of control packets is zero bytes. As Table 4.1 and Fig. 4.3 indicate, the fraction of control packets decreased after the link upgrade from 56.59% to 47.69% and then slightly increased in Long-term to 49.72%. The number of retransmissions follows a similar trend. This overall decrease in control overhead can be attributed to improved network performance, which resulted in less protocol overhead from retransmissions and repeated acknowledgements, as well as fewer attempts to re-establish failed TCP sessions. The uptick in retransmissions and control packets over the Long-term can be attributed to decrease in performance due to the increase in offered load to the new saturation point.

Temporal trends. We now discuss performance trends over time. We evaluate byte count in Fig. 4.4, which plots the average on a weekly basis aggregated per hour. For example, the first data point of Fig. 4.4 presents an average over all occurrences of the first hour of Monday, for each of the one month time windows. As the figure shows, there is a clear diurnal pattern in link utilization. Furthermore, the amount of traffic generated during weekdays differs from that on weekends, with weekday traffic loads typically being heavier. The increase in traffic after the network upgrade is also observable in the figures.

Table 4.2: TCP flow analysis.

Period	Total GBytes	Flow size (B)	IAT (s)
Pre-upgrade	105	3445	1.92
Post-upgrade	145	7708	1.49
Long-term	183	8103	1.91

4.2.3 TCP performance analysis

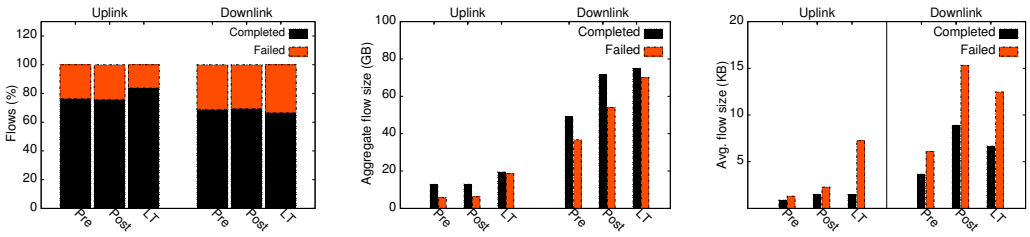
Our analysis shows that more than 93% of the traffic traversing the gateway in Macha is TCP. The performance of TCP improved significantly after the link upgrade. One factor indicative of this improvement is the *bytes in flight*, which is the fraction of sent data that has not yet been acknowledged. The bytes in flight is influenced by the TCP window size: the better the link performance, the larger the window size, which allows more data to be sent on the link before an acknowledgement is received. As indicated in Table 4.1, the TCP window size increased Post-upgrade and Long-term, allowing the amount of bytes in flight to ramp up. Fig. 4.5 presents a CDF of bytes in flight for the three periods. Immediately after the upgrade, the bytes in flight drastically increased and continued growing over the Long-term.

We continue our analysis by exploring TCP flow trends following this improved TCP performance. In order to extract uni-directional TCP flows from our `tshark` captures we develop a tool that examines packet signatures (sourceIP-sourcePORT-destinationIP-destinationPORT) and time-stamp and groups them in flows accordingly. We start by presenting general trends of TCP flows in Table 4.2. The total bytes associated with TCP flows increased after the upgrade and continued growing in Long-term. This increase in bytes is due to increased demand in browsing and streaming applications as well as increased rate of completion of larger TCP flows. We evaluate flow success and failure rates later in this section.

We next examine the average flow size across the three periods. As we can see in Table 4.2, the flow size doubled Post-upgrade and then continued increasing in Long-

Table 4.3: TCP flow uplink (UL) and downlink (DL) characteristics.

	Total GBytes		# of Flows (x10^5)		Pkt size (B)		Flow size (B)	
	UL	DL	UL	DL	UL	DL	UL	DL
Pre	18.65	85.9	189	194	132.7	616.0	988.9	4427
Post	19.26	125.7	114	116	158.6	877.0	1691	10856
LT	38.14	145	157	168	227.7	787.6	2422	8613



(a) Percent of flows (b) Aggregate bytes (c) Average flow size

Figure 4.6: TCP flow success and failure in uplink and downlink direction.

term. The increase of flow size can be attributed to different applications utilizing the link immediately after the upgrade and in Long-term. Indeed, we see many software updates Post-upgrade, which are then replaced by other applications as we explore in section 4.2.4. The average packet inter-arrival time (IAT) decreased Post-upgrade and then increased Long-term.

Uplink and downlink flows. Next we differentiate flows into uplink and downlink to analyze direction-specific trends. In Table 4.3 we first present aggregate bytes in each direction. Both uplink and downlink bytes increased after the link upgrade. While downlink increased rapidly, uplink remained almost unchanged Post-upgrade but then increased drastically over the Long-term. Average uplink packet size and flow size increased Post-upgrade and in Long-term. At the same time, downlink packet and flow sizes increased Post-upgrade and then slightly decreased over the Long-term. These trends can be explained with differences in applications accessing the Internet, as well as with changes in network performance due to link saturation in Long-term. The rapid increase in downlink activity Post-upgrade is due to an increase in automated activities

such as software updates. The increase in uplink happens more gradually, which is attributed to a slower change in user behavior and, in particular, a gradual increase in content upload attempts.

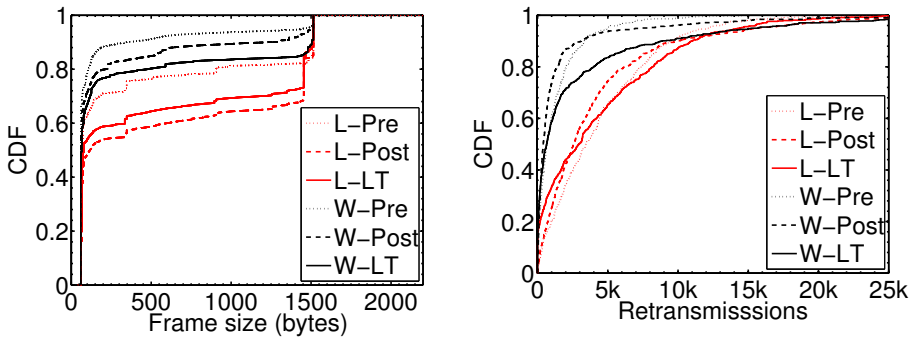
Finally, we concentrate on the number of flows. As we can see in Table 4.3, the number of flows in both up- and downlink directions decreased dramatically Post-upgrade and then increased. The initial decrease can be attributed to a higher rate of successful flow completions, which directly results in fewer flow re-initializations. The subsequent increase in the Long-term is due to a combination of increased user activity as well as an increase in flow failure rate as user demand again reaches link capacity.

TCP flows success and failure. We now focus on flow completion and failure. In compliance with RFC 793 that mandates the operation of the TCP protocol, we accept that a FIN packet indicates a completed flow, while lack of a FIN packet or exchange of a RST (reset) packet indicates a failed flow. Fig. 4.6(a) presents the fraction of completed and failed flows in uplink and downlink in each period. The completion rate of uplink flows remained unchanged Post-upgrade and then slightly increased in Long-term. On the other hand, the downlink flow completion rate remained unchanged. In Fig. 4.6 we also analyze success and failure trends correlated with byte volume and flow size. Fig. 4.6(b) plots the aggregate flow size in bytes for each direction. The aggregate size of both completed and failed uplink flows remained the same Post-upgrade and then increased in the Long-term. Unfortunately, the amount of bytes in failed flows approaches the amount of bytes in completed flows, which indicates that, while users were likely more successful in uploading content, over the Long-term half of the total content that users generated failed to upload. Similarly, in terms of total size of downlink flows, we see a gradual increase in successful downloads; however, over the Long-term the aggregate size of download flows that failed also increased, nearly reaching the aggregate size of successful downloads.

We evaluate average flow size of completed and failed flows in Fig. 4.6(c). We find the size of an individual flow by summing the packet sizes of all packets associated with the flow. In the uplink direction, the average size of failed flows over the Long-term is four times larger than the size of completed flows. This implies that smaller content uploads such as Facebook posts and small images are more likely to succeed, while larger uploads of videos or high quality pictures had a higher probability of failure. Similarly, the average size of failed downlink flows is persistently higher than that of completed flows. This points to the success of smaller flows, such as e-mail and web access, while the increase in downlink average flow size for failed flows is likely due to increased attempts to download larger files, such as video content.

Windows vs. Linux. Lastly, we evaluate the TCP performance of two of the most prevalent operating systems used in Macha: Windows and Linux. Using the observed TTL values, we were able to distinguish between the two operating systems [140]. Linux implements CUBIC TCP, which has optimized congestion control mechanisms for high bandwidth networks with high latency. This optimization occurs by calculating the window size according to the last congestion event. In this way, CUBIC TCP measures congestion independently from long RTTs [60]. This differs from Windows, which implements TCP Reno in Windows XP and Compound TCP in Windows Vista and subsequent Windows versions. TCP Reno and Compound TCP base window size on the RTT – window size increases with low RTT values and decreases with high RTT values. This method of congestion calculation causes Windows machines to interpret network latency as network congestion [101].

In Fig. 4.7 we plot frame size and retransmissions per hour for Windows (W) and Linux (L) machines. As we can see, Linux maintains higher mean frame size over all three periods. Thus, it is much more aggressive in pushing data onto the link. Naturally, this results in more retransmissions per hour in comparison with Windows. Linux's



(a) Frame Size (b) Number of Retransmits

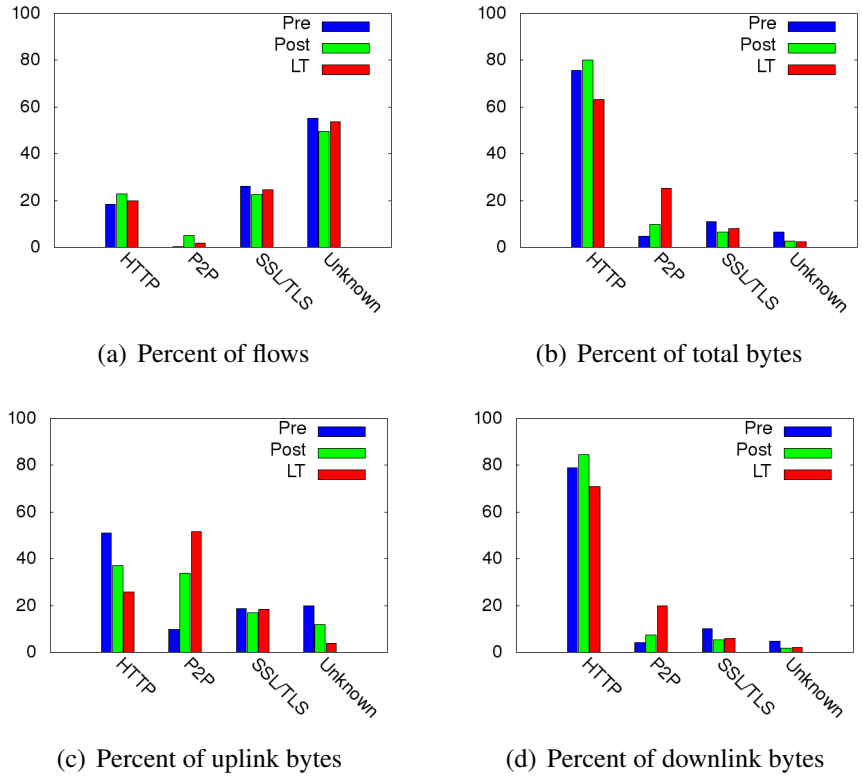
Figure 4.7: Comparison of TCP performance in Windows and Linux.

aggressive behavior, however, leads to higher achieved throughput of 487.6 Kbps in comparison with Windows, which only achieves 106.2 Kbps in the Long-term.

Most popular services. We analyze the most popular services accessed by Machans in the three periods. For this analysis we make use of a tool called Tstat[†], which can perform layer-7 packet inspection to determine service type.

Our initial results show that the most popular services across periods were HTTP, P2P and SSL/TLS. Thus, the remainder of this section focuses on those services. Fig. 4.8 presents our results for (a) percent of bi-directional flows to each service, (b) percent of total bytes, (c) uplink bytes and (d) downlink bytes. We examine trends across the three periods and look for correlations between different services in order to capture changes in user behaviour. A substantial amount of the flows could not be classified by the layer-7 inspection and were labelled as Unknown. As we can see on Fig. 4.8(a) and 4.8(b) this Unknown traffic constituted the majority of the flows observed across all periods; however, the fraction of bytes due to Unknown flows is minimal. We postulate that the Unknown traffic was due to malware such as port scans, which generated a large volume of small flows. The fraction of bytes due to Unknown traffic (Fig. 4.8(b), 4.8(c) and 4.8(d)) decreased Post-upgrade and over the Long-term. This is likely due to

[†]http://tstat.tlc.polito.it/



(a) Percent of flows

(b) Percent of total bytes

(c) Percent of uplink bytes

(d) Percent of downlink bytes

Figure 4.8: Tstat analysis of service types.

the successful completion of software updates, which allowed computers to better defend against malware. SSL/TLS is the next most accessed service, followed by HTTP. On the other hand, in terms of generated bytes, HTTP is much more prevalent than SSL/TLS. Lastly, P2P is the least popular in terms of percent of flows; however the bytes due to P2P flows are substantial.

Exploring trends across the three analyzed periods we see a reverse correlation between the bytes due to HTTP and P2P flows over the Long-term (Fig. 4.8(b), 4.8(c) and 4.8(d)). As Fig. 4.8(b) shows, there was an increase in both services Post-upgrade. Over the Long-term, the bytes due to HTTP activity dropped and those due to P2P flows increased significantly. This indicates a shift in user interest from web browsing to P2P file downloads. Further analysis of the upload and download bytes (Fig. 4.8(c)

and 4.8(d)) confirms this trend. We see a gradual increase Post-upgrade in both uplink and downlink P2P bytes, while a more substantial increase can be observed over the Long-term. Facebook, Google and software updates were among the top applications accessed through HTTP. 40% of the P2P traffic was through BitTorent applications and the remainder was other unclassified P2P traffic. The nature of BitTorrent applications provides a partial explanation of the increase in both uplink and downlink P2P bytes. Particularly, when a user downloads a torrent, this user is simultaneously a seeder; i.e., this user becomes a source of the file and uploads that file to other torrent clients. As a result, 28% of all the uploads were BitTorrent uploads, or in other words, (potentially unintentional) seeding activity. The remaining 72% of the uploads were user-initiated and consisted of HTTP (23%), SSL/TLS (19%), unclassified P2P (23%) and other (7%).

4.2.4 Network usage

The most prevalent application protocol used in Macha is web-browsing. 87% of the Pre-upgrade traffic in up- and downlink direction was a combination of HTTP and HTTPS. This number remained almost unchanged Post-upgrade, and dropped to 71% in the Long-term. At the same time, P2P and Unknown traffic, which includes services to unspecified ports (e.g. Skype and BitTorrent) increased in the Long-term. This is a strong indication of a shift of usage habits to more real-time services, which is typical for well-connected Internet users.

In this section we investigate web traffic to understand user behavior. We correlate our findings about popular applications with network performance and make inferences about the user experience based on this correlation.

Popular URIs. We begin our analysis by evaluating popular web services. Fig. 4.9 shows web URI requests classified by the destination domains and includes the top 14

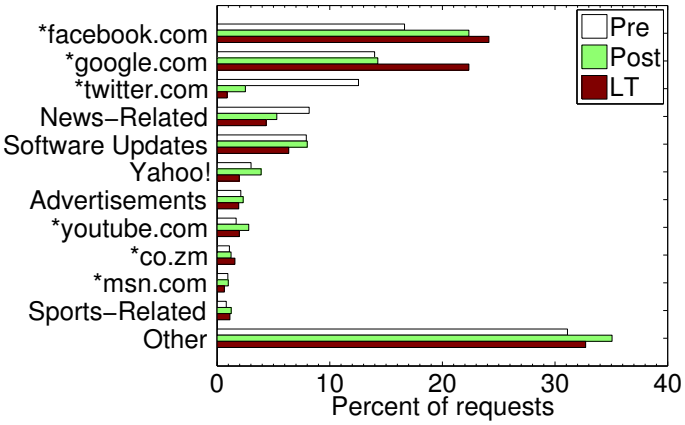
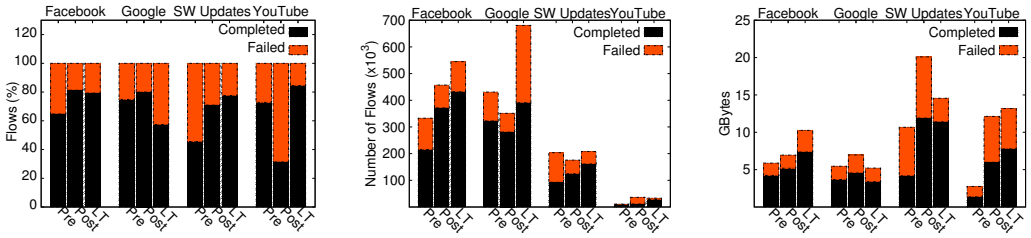


Figure 4.9: Popular URI Requests.

requested sites. For clarity of presentation we combine related sites (e.g. Facebook with the associated Content Delivery Networks). Facebook and Google are clearly the most popular sites. Both sites see a significant increase in the percentage of requests after the link upgrade, further extending their dominance. At the same time, access to Twitter, the third most popular domain Pre-upgrade, dropped significantly. The "News" classification includes *postzambia.com, *lusakatimes.com, and BBC news sites. The popularity of these websites is important as it shows user interest in local content, a pattern also seen in [75].

Software update sites such as those associated with Windows, Adobe, and Ubuntu remain relatively unchanged throughout the measurements; however, as shown later in this section, their completion rate significantly increased Post-upgrade. While requests for multimedia-rich sites or large binary downloads remained the same across periods, the actual traffic associated with such requests increased as more requests were successfully completed. We explore TCP sessions count and size later in this section.

Advertisement-related sites are the seventh most popular request type, representing roughly 2% of all requests. Traffic generated by such requests is equivalent to wasted bandwidth as most advertisements are targeted at more affluent urban consumers and



(a) Percent of flows (b) Number of flows (c) Aggregate bytes

Figure 4.10: TCP flow success and failure for URIs of interest.

are likely of no interest to users in rural Zambia. As bandwidth is clearly a scarce resource in this network, such wasteful access to advertisements can lead to further deterioration of user experience.

Following our URI findings we evaluate TCP flow patterns associated with four of the most accessed online services: Facebook, Google, YouTube and Software updates. For this analysis we combine the previously extracted uni-directional flows into bi-directional sessions based on flow signature and timestamp. We then extract flows of interest based on the URIs that have been accessed in the corresponding session. Fig. 4.10 plots (a) the percentage and (b) the number of flows as well as (c) the aggregate bytes over each period for the four services. The results are divided in terms of flow completion and failure. Both the number and total bytes associated with Facebook flows increased over the three periods. This trend is different than the one followed by Google, which in terms of number of flows remained almost unchanged Post-upgrade, but increased over the Long-term. Similar to Facebook, YouTube also increased immediately after the upgrade both in terms of flows and aggregate bytes. Interestingly, the failure rate of YouTube flows was high Post-upgrade and then decreased. This might be due to software updates using a large fraction of the bandwidth Post-upgrade, which caused YouTube to fail more often. Of note, while only 16% of YouTube flows failed in the Long-term, those accounted for 40% of the YouTube flow

bytes. This implies that large flows were most often the ones to fail. Due to the increased interest in access to real-time streaming services such as YouTube, the network quickly achieved its maximum capacity, inhibiting these services with substantial flow failures.

Lastly, we look at TCP flows from software updates. The number of such flows decreased slightly Post-upgrade and then increased in the Long-term. Our analysis indicates that the short-term decrease is due to improved network performance resulting in fewer TCP session re-initializations. Furthermore, the quantity of bytes associated with software updates doubled immediately after the link upgrade. This is likely due to long-postponed software updates finally being able to complete. We see a decrease in software update bytes in the Long-term due to successful completion of updates in the period Post-upgrade.

Table 4.4: HTTP Response Codes

Response	Pre-upgrade	Post-upgrade	Long-Term
200	4,289,578	3,333,240	4,667,380
400	5,933,008	2,627,842	3,514,872
408	17,146	68	162
Total	12,638,744	7,507,975	10,186,110

Next, we measure HTTP response codes in an effort to find discernible differences between observation periods. We find noticeable changes in three response types: 200, 400, and 408. 200 (OK) responses indicate a valid request for which an HTTP server can correctly craft a response. As shown in Table 4.4 the percentage of HTTP 200 responses increases more than 10% after the link upgrade. 400 (Bad Request) errors indicate a request that the web server does not understand. These errors typically are caused by bad syntax or potentially a host infected with malware that sends poorly defined HTTP requests. The table shows 400 errors decrease significantly after the link upgrade. We believe that this could be due to two changes. First, immediately after the upgrade hosts could have implemented overdue software updates which could rectify

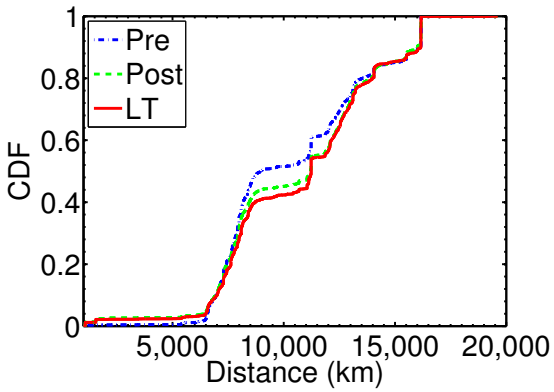


Figure 4.11: Flow distance from Macha; CDF.

browser version issues associated with request format. Secondly, in a similar fashion to operating system software, anti-virus software was updated to newer versions which could potentially allow for the detection and removal of malware on hosts. The final response code we investigate is 408 (Request Timeout) which indicates that the server was expecting a request from the client in some amount of time and the client failed to produce such a request. Such errors occur in networks with very limited bandwidth or where multiple packets are dropped some point along the path. The number of 408 errors decreases dramatically after the link upgrade. This is an encouraging result, as it shows that even a small bandwidth increase can make a large difference in the user experience.

4.2.5 Flow Distance

We investigate the network traffic using geographical information to characterize usage. For each traffic flow we identify the external node IP address. Using these IP addresses, we query the MaxMind GeoIP database [5] to correlate each flow with geographic coordinate information. Our preliminary investigation involves calculating the straight-line distance between Macha and the given coordinates for the other side

of each connection using the Haversine formula [129]. Fig. 4.11 shows the CDF of the flow distances from Macha in each of the three observation periods. We find that flows generally occur over longer distances in the periods after the network upgrade. Of note is the large increase between the Pre period and the Post period in the roughly 8,000 to 12,000 km range. While Long-Term flows show even longer distances as compared to Post-upgrade, the increase is not as pronounced. We posit that a potential reason for the increase in distances from Macha is the result of a better user-experience after the network upgrade, which encouraged users to access such services that are physically further away.

We also use the GeoIP database to find the country code for each external node. We calculate the number of bytes associated with each country code and rank them. Interestingly, traffic to and from nodes in Zambia itself increased dramatically after the network upgrade. In the Pre period, Zambia ranks as the thirteenth most popular country in terms of bytes, representing 0.9% of all traffic. In the Post period, Zambia jumps up to rank second representing 23.4% of all bytes; in Long-Term it is ranked third with 12.1%.

Table 4.5: Measured radius of gyration.

Period	Distance (km)
Pre-upgrade	6,363.26
Post-upgrade	6,851.41
Long-term	7,096.86

Our initial distance findings lead us to investigate not only the distance from Macha that flows represent, but also the overall "worldliness" of the network flows. That is to say, we investigate the distribution of the geographic coordinates in order to further characterize network usage. We utilize the Radius of gyration metric to provide a value for the spread of the data. Radius of gyration has been used extensively to characterize user mobility in wireless networks [59] and provides a technique for measuring dis-

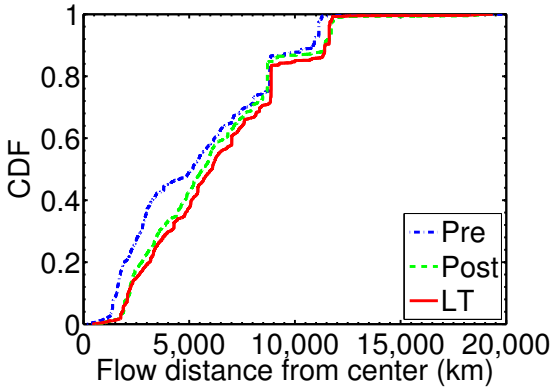


Figure 4.12: Flow distance from center mass; CDF.

persion. It can be understood as the range of observed points up to time t and can be calculated by the formula:

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\vec{r}_i^{\,a} - \vec{r}_{cm}^{\,a})^2}$$

where $\vec{r}_i^{\,a}$ represents the i^{th} coordinate for period a and $\vec{r}_{cm}^{\,a}$ is the calculated center-mass for the period. $n_c^a(t)$ represents the number of points measured up to time t. Table 4.5 shows the results for the observation periods. We see that each successive period shows an increase in the radius compared to the prior periods. This means that not only are flows connecting to locations further away from Macha as seen in Fig. 4.11, they are actually spreading out further as well. Assuming users are behind the majority of network traffic we can argue that network users are connecting to content from a larger geographic variety of the world.

We verify these results in two ways. First, we find the distances between each flow and the center-mass and plot the CDF shown in Fig. 4.12. As expected, we see longer-distance flows in the periods after the network upgrade. Further, the increases seem to be incremental and uniform rather than drastic changes. We also investigate the center-mass values to determine whether the upstream provider (which changed along with

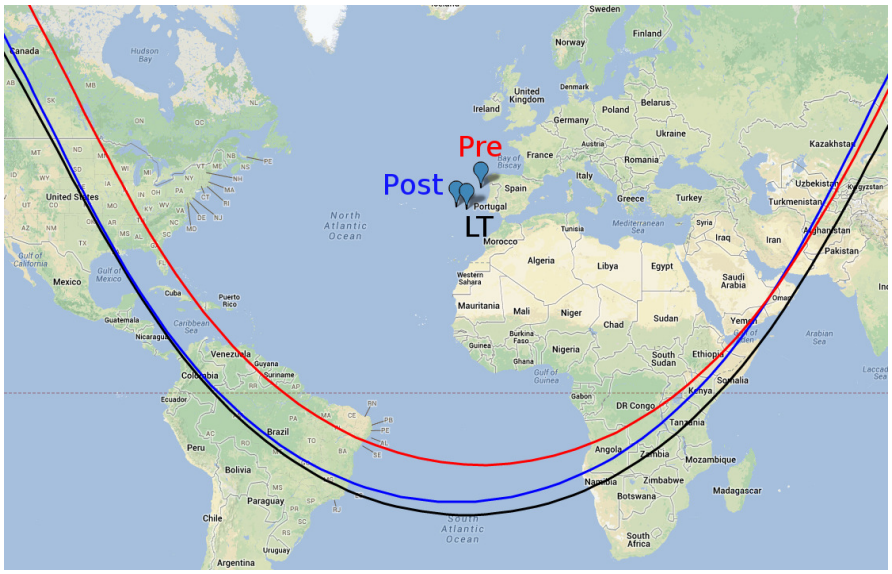


Figure 4.13: Center-mass points with radii of gyration.

the network upgrade) drastically altered the distribution of external nodes. While we expect the center mass values to be different for each period, we also expect them to be somewhat clustered. Should the upstream providers apply unexpected policy-based traffic routing (e.g. resolving all CDN queries to a particular location), we expect to see the center mass values vary dramatically between the Pre period and those after the upgrade. Fig. 4.13 shows the center-mass points for each period as well as each calculated radius of gyration. We find the three center-mass values are within a reasonable range of each other given the global scale and as such we do not credit the upstream provider with the radius of gyration increase. Given these results we are confident that the increase in spread can be credited to an increased geographic diversity of external nodes.

4.2.6 Long Term Trends Persistence

To verify whether the observed long term trends persist over time, we examine network performance and utilization in one additional time interval seven months post-upgrade. In order to assure that a change in user population would not skew our results, we focus on November, since this month has roughly the same number of users as the previous months of interest. We measure the same set of characteristics as for the previous three months and compare results from November and July (our Long-term month). Our results show that the overall network utilization in November was similar to that of July. The user behavior, however, changed, whereby users were no longer attempting to upload large files to the Internet but instead increased their download activity. As a result, the total uploaded traffic in November decreased and became comparable to that before the upgrade, while the total downloaded bytes increased significantly by 60

We analyze a total of two and a half weeks of November. The reason for not analyzing the full month is two-fold. First, the first four days of the month were not captured in our traces due to failure of the managed switch that was mirroring the network traffic for monitoring purposes. Second, the Android phone phenomenon that we observed in July was aggravated throughout November, whereby approximately 4 million out of a total of 5 million requests sent to Google were produced by this phone. In order to conduct an objective comparison of network performance and utilization, we discard the minute intervals throughout November where this phone was present in the network, as the presence of the phone significantly skews the results. For the remainder of this section we present results from the two and a half weeks analyzed; we refer to this traffic as the November traffic or simply November.

In November, a total of 158 GB traversed the network, which consisted of 373 million packets. This is about 1GB/day more than what we observed in July. This increase in the total amount of data that traversed the network is due to a combina-

tion of improved TCP performance and larger download attempts. The improved TCP performance is indicated by factors such as control overhead and bytes in flight. The control overhead in November decreased to 45.2% (compared to 49.7% in July) and the bytes in flight increased from 10.5kB to 36.4kB. We postulate that the large value of bytes in flight is due to longer TCP flows, which allows enough time for TCP to ramp up the amount of bytes sent and achieve higher throughput. The latter results in increased bytes in flight. We also examine the possibility that this aggressive behavior is due to more Linux traffic traversing the network. Our verification, however, shows that the ratio of Linux and Windows packets that traversed the network is comparable to that in July. Thus the reason for this aggressive behavior is not related to increase in the amount of Linux traffic. Lastly, we examine retransmissions as an indicator of TCP performance. We observe an increase of retransmissions from 1.16% in July to 3.68% in November. While this increase contradicts the TCP performance improvement, a closer look at retransmissions over time shows that there were several short periods where the network was experiencing poor performance, which resulted in few tens of flows with millions of retransmissions. If we omit those flows, the retransmission performance is comparable to that in July.

We now examine uplink and downlink trends in an effort to assess whether user experience in sharing and accessing content persisted throughout November. The total uplink bytes in the November traffic is 11.7GB, which is approximately 650MB/day. This is a decrease in the uploaded traffic in comparison with July, where uploads were 1.3GB/day. In fact, the amount of uploaded bytes per day in November is comparable to that before the upgrade (620MB/day). As we saw in section 3.3 the performance of uplink flows deteriorated dramatically as users began to upload more content. Thus, the decreased amount of uploaded bytes in November can be explained by users having become discouraged from sharing their content online by the poor upload performance.

Table 4.6: TCP flow uplink (UL) and downlink (DL) characteristics.

	Percent of Flows		Aggregate GB/day		Avg. Flow Size (kB)	
	July	November	July	November	July	November
Failed DL	38	42	2.33	2.08	12	6.2
Success DL	62	58	2.5	4.21	6.5	9
Failed UL	19	21	0.633	0.147	7	0.9
Success UL	81	79	0.667	0.504	1.5	0.8

At the same time the amount of downloaded traffic increased from 4.8 GB/day in July to 6.3 GB/day in November, indicating a clear trend of more Internet traffic consumption than generation.

To get a better understanding of how success and failure of uploads and downloads influences users inclination to share and consume content, we separate the November traffic into completed and failed TCP flows and analyze these flows in the uplink and downlink directions. As Table 4.6 shows, there was a slight increase in the percentage of successful uplink and downlink flows. At the same time, the traffic volume measured in aggregate bytes changed more substantially in both directions. The aggregate bytes in failed uploads decreased by 77%, reflecting on peoples decrease in interest to upload content online. Similarly, the aggregate bytes of successful uploads decreased from 667MB to 504MB/day. In line with this trend, the average flow size of both completed and failed uplink flows decreased in November. The aggregate of successfully download bytes increased by 60% in comparison with July, indicating a rapid increase in the amount of consumed traffic.

To better understand this shift in usage characterized by rapid decrease in uploads and growth in downloads, we examine the most popular layer-7 services in November. We observe a notable change in usage as compared with July, whereby P2P traffic decreased and HTTP increased. The fraction of P2P bytes in November dropped to 9%, in comparison with 25% in July. This decrease was shared between uploads and downloads but was much more pronounced in the upload direction: P2P bytes dropped

from 52% to 7%. Because the P2P uploads constituted a large portion of the overall uploads in the network (51%) in July, their decrease had a significant impact on the overall upload volume observed in November. Simultaneously, the percentage of HTTP downloads increased from 63% to 78%.

The goal of the analysis presented in this section was to evaluate persistence of network performance and usage in the post-upgrade period. While network performance improved in comparison with July, the usage of network services changed. In particular, we observe a rapid decrease in the volume of upload flows, while more emphasis was put into downloads. We saw a shift of usage from HTTP immediately after the upgrade to P2P file transfers in July and back to HTTP in November, which indicates that users were eager to try new services but ultimately reverted back to HTTP, likely due to poor user experience.

4.2.7 Benchmark

In this section we answer the question "Given the limited gateway capacity in Macha, how well can the network perform?". To this end we provide a benchmark of network performance in November and compare this benchmark performance with the actual observed network performance. Intuitively, the performance of services such as Facebook and Google is influenced by the specific service configuration. For example, whether services are centralized or use content distribution networks, and they handle poor network performance, can directly influence the user experience. In order to be able to objectively compare the performance of the Macha network with a benchmark, we need to ensure that the benchmark setup resembles the same service configuration as the one used by Machans during the trace capture period. At the same time, service configurations often change to meet the ever growing user demand and application growth. Thus, we base our benchmark on the November traces we capture

in Macha. In particular, we analyze parts of the trace when fewer users used the network in order to ensure there was minimal contention on the link. Our analysis focuses on TCP performance and shows that the benchmark traffic performs significantly better than the average performance in November.

Therefore, to generate the benchmark trace we divide the November capture into one minute chunks and consider the minutes with three or fewer users. This method generated a total of 2853 minutes (nearly 48 hours), which we use as our benchmark. The website access distribution during those hours is comparable to that during busy periods. The benchmark traffic amounted to 5.88GB. Out of a total of 9,556,817 packets, 16% were control packets, which is a significant reduction in the amount of protocol overhead as compared to the more general scenario. The average RTT decreased from 157ms to 119ms. Because the same websites were accessed, this likely indicates that there was less queuing delay, as the ISP gateway and/or core network was less overloaded.

With high-delay, low-speed connections, rural users are often limited in their ability to utilize the Internet. Furthermore, the benchmark presented in this section clearly shows that when multiple users share a limited Internet connection, the end user experience can be further deteriorated. In less overloaded scenarios, however, the TCP performance becomes comparable to that of the western world, whereby the RTT reduces and the amount of control overhead drops significantly. Thus, if a single 2MBps link such as the one in Macha was used by several users, there is a hope that the Internet experience can be on par with that in the western world. When a single 2MBps connection is shared between tens or hundreds of users, however, the network performance will continue to negatively impact the user experience.

4.3 Related Work

Our work builds upon earlier analysis of Internet access in Macha [78]. This prior work focuses on network performance and usage during a two week measurement period in 2010. Other work analyzing rural Internet usage patterns includes [50] and [69]. Web traffic from Internet cafés and kiosks in Cambodia and Ghana is analyzed in [50]. Here the focus is on the characterization of HTTP traffic to guide caching techniques for web users in developing regions. Ihm et al. [69] focus on understanding the network traffic in developing regions as compared to their OECD counterparts. This paper characterizes national traffic patterns based on network usage, with the goal of improving caching techniques for developing regions. Anokwa et al. identify the impact of latency on network performance in developing regions and propose a flow-based prioritization scheme as a solution [23]. In contrast, our work focuses on a smaller scale and characterizes web traffic in order to ascertain the impact of a network upgrade on usage and performance.

Our analysis of TCP performance builds upon the measurements employed by Johnson et al. [78] by engaging a more in-depth analysis of TCP performance, including measurements of TCP windows, retransmissions, inter-packet arrival times, RTT, and packet sizes. Performance of CUBIC TCP, Compound TCP, and TCP Reno interactions is measured via simulation of high delay wireless networks in [15]. This work has an explicit interest in measuring goodput and TCP fairness. Our analysis is based on TCP performance in a low-bandwidth, high-latency real network; we measure TCP performance in aggregate and as separated by operating system network stacks, and draw conclusions about TCP fairness of different variants found on the network.

4.4 Next Steps

Our analysis of a network upgrade in a rural community indicates that even a small increase in access bandwidth can improve the usability of the network: for example, successful software updates and updated anti-virus protection immediately after the upgrade grant better performance in HTTP request generation and, overall, decrease the traffic due to malware activity, resulting in possibilities for better performance. While these results are encouraging, incremental increase of available bandwidth can often bring only marginal improvement of user experience, as indicated by the large volume of failed requests in the case of Macha. In the face of such increased usability but still low quality of user experience, the need of systems such as VillageShare [79] (and others [50,73,144]), that can intelligently manage activities in the network, is even more pronounced.

One immediate need that arises from our analysis is the one of prioritizing bandwidth allocation to critical services. For example, as usage patterns in Macha did not immediately change post-upgrade, critical software updates were finally able to complete. This, in turn, resulted in rapid improvement in browsing experience (as indicated by the drop in HTTP Bad Request and HTTP Request Timeout messages in the network) on one hand and by the decrease of traffic associated with malware on another. This observation hints of a need for a system that can detect critical services and emulate a bandwidth increase for such services.

Such a system would be able to perform real-time detection of network traffic anomalies (for example increase in abnormal HTTP requests or traffic to ports associated with viruses) and would prioritize bandwidth assignment for software updates. Two major concerns arise with regards to such a system. First, to ensure that such bandwidth prioritization does not compromise the user experience, this functionality can be embedded in time-shifted proxies such as [50, 79, 144]. Time shifting to off-

peak hours, however, runs the risk of users turning off their computers, which brings us to the second challenge in such system design. To handle this, local caching techniques [73, 79, 144] can be employed which make particular content (e.g. software updates) available in the local network for use during peak hours.

4.5 Discussion and Conclusion

We utilize a unique dataset from a rural sub-Saharan village that captures usage before and after an Internet access speed upgrade. We study the effects of this upgrade on the network performance and user behavior. We find that performance improved immediately after the upgrade, whereby automatic services that were previously failing due to slow access speed were finally able to complete. With improved network performance, subscribers were encouraged to use more bandwidth-demanding services such as YouTube video streaming. There also was a substantial increase in attempts for content sharing online, whereby the uplink byte volume doubled in the Long-term. Unfortunately, with the increase of upload attempts, the failure rate of uploads grew as well.

Another trend that stood out from our analysis is the stark difference in performance between Windows and Linux operating systems. As our results show, Linux outperforms Windows, achieving five times better throughput on average in the Long-term. This makes a case for careful selection of operating system and/or modifications of the network stack to facilitate better networking performance in bandwidth-constrained environments.

Internet access upgrade in the context of developing rural regions is not a trivial task. Although such upgrades are perceived to lead to overall better performance and user experience, this is not always the case for communities that are largely bandwidth-

impaired. In such communities, an Internet upgrade can be just a small increment to the more substantial access speed that is needed to accomodate modern web content and applications. Each such increment gives users the ability to more fully utilize the modern Internet with bandwidth-intensive applications; however, it is clear that in the developing regions case, even an eight-fold increase in network capacity can be not nearly enough. Many rural communities like Macha have a long way to go before their Internet experience parallels that of users in the Western world.

4.6 Acknowledgements

The material in this chapter is based on joint work with Paul Schmitt, Morgan Vigil and Elizabeth Belding.

Chapter 5

Sociological Approach to Identifying Challenges

5.1 Introduction

Voice technologies are revolutionizing communication in remote areas. For example, access to cellphones in Macha allowed farmers to overcome what they call the "briefcase buyers problem", whereby businessmen who buy maize come to the village and often attempt to exploit farmers who are unaware of market prices, buying crops at extremely low prices. Farmers in Macha can now call the Zambian Food Reserve Agency (FRA) to get information about crops prices. Furthermore, a new communication structure of a combination of Internet and cellphones has emerged in Macha, whereby a farmer who has access to both the Internet and a cellphone can check crop prices online and send text messages to fellow farmers.

One domain that have benefited significantly from technological innovation is rural healthcare. Our work focuses on immunization distribution, which have seen an increased effort in incorporating Information and Communication Technology (ICT) for improved vaccination distribution to children and infants [41, 81]. Existing systems, however, are often designed according to western models and operate under assump-

tions that do not hold in developing countries. To bridge the gap between western perceptions and the reality in rural areas we work towards better understanding of the ICT needs in immunization distribution. To this end we volunteer at the hospital in Macha and conduct extensive interviews with local residents to (i) better understand the state of rural telephony and (ii) assess the applicability of cellphone technology to healthcare.

5.2 Voice communication in rural Zambia

During our field work in Zambia, we visited Macha and three surrounding villages – Hamoonde, Chikanta and Mapanza. All of these villages use VoIP, cellular technology or both as a means for voice communication. Within the four communities the level of availability of these technologies varies widely, resulting in fundamentally different voice communications usage patterns. We summarize the current cellphone coverage and usage in these communities below.

Macha. Macha is a village in the Southern province of Zambia, located 70km from the closest city, Choma. There are 135,000 people in the area spread over a large radius of 35km. The village is connected to the national power grid; however, individual households rarely have access to electricity. This sparse availability of power results in creative solutions, such as travel to powered public spaces to charge cellphones; others charge their phones while at work. However, while power is technically available, the reality is that the availability and quality can vary on an hourly or daily basis.

Cellphone coverage in Macha was first introduced in 2006 by Celtel (now Airtel). By 2012, MTN was a second active cellphone provider in the village. Residents with a cellphone subscription can use plain voice and text messaging and, where available, low data rate GPRS service for Internet access. The coverage provided by the two operators

is largely available in the central part of the village; coverage is inconsistent and spotty in residential areas. A large percent of the people we interviewed had subscriptions with both cellular providers to increase the likelihood they have cellphone coverage at any given time. Our first hand and anecdotal experience, however, indicates that the failures of both commercial networks are highly correlated and often coincide with power failures.

In terms of VoIP, Macha is connected to the Internet through one of Zambia's Internet service providers. A monthly subscription limited at 1GB traffic costs 30 USD. Together with electrical unavailability, this high price renders home Internet access impossible for most residents. A small subset, however, can access the Internet from work or from an Internet café in order to use VoIP services.

Hamoonde. In contrast with Macha, neighboring Hamoonde, only 5km away, has very sparse cellular coverage, no Internet access and no connection to the power grid. Villagers identified the lack of electricity as the main obstacle to computer and Internet access in the village. People we spoke with, however, owned cellphones and expressed their desire to have more widespread cellular coverage. Residents carry their phones for use when they enter an area where cellular access is available. We encountered a case where a cellphone was used as a voice gateway to the Internet, whereby one resident of the village regularly called a relative in the city with access to the Internet to ask her to lookup information online.

Chikanta. Chikanta has population of about 40,000 and is 25km away from Macha. There is no access to electricity within the village; commercial cellphone coverage was introduced in some areas of the village in the beginning of 2012. A commute of 2km is common for village residents to reach an area with cellphone coverage. Interestingly, Chikanta has been largely influenced by the technological innovations within Macha. In 2010, two years before commercial cellular coverage was available in the village,

Chikanta obtained an Internet container kiosk* that connects the village to the Internet through a satellite link and that runs on solar power. This allowed people from the village to start gaining experience with VoIP before they had easy access to cellular technology.

Mapanza. Mapanza is 7km away from Macha and has population of about 2000 people. There is cellphone coverage in the central part of the village. Similar to Chikanta, Mapanza recently obtained an Internet container kiosk that operates as an Internet café and uses a slow satellite link to connect to the Internet.

5.3 Social Surveys

During our three week field trip to Zambia in June/July 2012, we conducted extensive interviews with local residents to better understand their usage and experience with voice communication technologies. Each interview was conducted privately with one interviewee, one interviewer and a translator if necessary. Part of the interview questions were of closed form (pre-selected answer options) while others were open-ended to enable free discussion of voice communication topics. Interview participation was voluntary and no material incentive was associated with participation. While our method did not produce a random sample of interviewees, we made an effort to interview people across age, gender, occupation and income groups.

In the course of three weeks, we interviewed 26 people total from Macha, Hamoonde and Chikanta. Our interviewees were between 20 and 48 years old; 14 were female and 12 were male.

*http://www.machaworks.org/en/projectwizard.html/project/25

5.3.1 Cellphone usage

All interviewees had owned a cellphone at some point; 96% owned a phone at the time of the interview. Those who did not wanted to recover access but did not have money to do so. People who owned a phone either bought the phone themselves (76%) or obtained it as a gift. Despite the high availability of cellphones among the group we interviewed, cellphones are very expensive relative to income; 68% of phone owners who bought their phone had to save between one month and one year to afford to buy that phone. Our interviewees shared that apart from buying the cellphone, the recurring cost for air time significantly adds to the financial burden for adoption of cellphone technology.

44% of our interviewees had more than one SIM card and most often identified one of two reasons for their subscription with multiple cellular networks. The first reason is associated with cost; inter-provider services cost more in comparison to intra-provider services. Thus, users often carry more than one SIM card and pick the one to use based on the network to which the person they are calling is subscribed. The second most popular reason for people having multiple SIM cards is reliability. As the quality of service often varies over time and space, users carry multiple SIM cards to increase the chance they will be able to use a cellphone network at any time. Despite this effort, a commute of 1-2km is typical for residents of Hamoonde and Chikanta to reach to an area with coverage and conduct a call.

When asked what is enjoyable about cellphone usage, 100% of our interviewees identified the opportunity to stay in touch with relatives, including those who live in other places. 80% indicated that they use the phone to obtain information, such as crop pricing for farmers, availability of goods for local businessmen who buy from the city and communication with regional authorities. Users with more advanced phones and data subscriptions identify the importance of staying connected to the Internet at

all times. With the low rate GPRS connection in Macha, however, all users with cell phone Internet access indicated that browsing is slow and frustrating.

5.3.2 Cellphones in support of immunizations

We start our analysis by outlining several common assumptions in western designs of systems for immunization distribution in rural developing regions. We then detail our findings from the in-person interviews.

One typical assumption is that personal immunization history can be kept easily by the health workers as they perform vaccinations. Our work with the immunization staff from the Mission Hospital at Macha, Zambia, unveils that immunization clinics are often understaffed and keeping personal records along with the day-to-day immunization routines is a very time consuming and challenging task.

Another common assumption is that digitalizing immunization records would help significantly improve the efficiency of storing and accessing patients data. Technology including such that facilitates digital data entry is just being introduced in the developing world. Thus, otherwise skilful health care staff often lacks basic technical skills such as typing. In fact, our work in Macha shows that at best people can hunt and peck and these are typically people involved with IT, not the health workers. As a result, digital data entry could turn into an additional hurdle that needs to be tackled by the already overworked health care staff. Furthermore, while adopting digital data entry, the immunization clinics still need to maintain their traditional practices of paper data entry, as this paper work is required on regular basis by the government officials. This results in doubling the effort put into keeping immunization history.

Challenges in distribution of vaccines are not only related to data entry and personal immunization history but also to outreach, especially to parents with children under the age of five, who do not attend school yet. Traditionally, information about vaccination

schedules and availability is distributed through posters, word of mouth and in Macha, through the local community radio. Information is typically disseminated a few days or even weeks in advance, and as one of our interviewees from Macha shares, ”It is very easy to forget which day exactly was vaccination day and miss an immunization”. Thus, having a technology that is intuitive to people to help remind them for upcoming immunization events is of great importance.

The results from our interviews unveil that the benefit from immunizations is very well understood in the local community. All the people interviewed had been immunized when they were young and some of them perceive immunization as a ”tradition” that needs to be followed, while others are keen to immunize their children, having experienced themselves the benefits of immunization in previous disease outbreaks, when they ”did not suffer from the corresponding disease” because they were immunized. Only one of the participants who had a child under her care, did not do all recommended immunizations; all other participants had immunized all their children against all recommended diseases.

Cellphone usage too is widely adopted by people from the community as well as health workers. 24 of the 25 interviewees owned at least one cellphone and SIM card and all the interviewees were very accustomed to using basic functionality such as call and text messaging. 24 of the participants were excited about the idea for using their cellphone to receive SMS reminders for upcoming immunizations for their children.

5.4 Conclusion

Our interview findings confirm the importance of cellphone technology for residents of remote areas. While the reasons for adoption of cellphone technology are not drastically different than these in the western world, the benefits for people in remote

communities without infrastructure and with intermittent income is much more pronounced. Obtaining information via cell phone, as opposed to in person after travel, saves both critical time and money. Unfortunately, while the benefits and need of cell-phone technology in remote areas are well understood, wide availability and affordability of such technology is still severely lacking.

Through our work with the hospital in Macha we have been able to identify the poor communication channels between healthcare personnel and patients as a major problem in vaccine distribution in rural areas. Future work in improved distribution of immunizations should leverage the opportunities presented by the availability of cellular networks for improved healthcare-patient communication.

Part II

Connectivity Solutions for Developing Regions

While wireless networks have lead the way in providing last hop access in more densely-populated areas, current network designs fall short in the case of rural developing regions. Current networks are (i) very infrastructure-intensive and (ii) rely on high-bandwidth gateway access links to connect to the Internet. Unfortunately, these two resources are often unavailable in the rural developing context, characterized with poor infrastructure and limited connectivity. Our usage analysis has shown that as a result rural areas are under-provisioned [154] in terms of cellular connectivity and suffer poor Internet access performance despite increases in available bandwidth [156]. At the same time, rural network usage exhibits specific characteristics: for example, 70% of the calls in Ivory Coast occurred under the vicinity of the same antenna. Such characteristics open space for a rethink of current designs in order to improve user experience.

We leverage this knowledge to design and build connectivity solutions with the goal of widespread adoption. While the focus of our network studies and system deployments is in rural sub-Saharan Africa, our networking designs are well-suited for a broad spectrum of infrastructure-challenged environments. We design solutions both for cellular access to ensure rapid adoption, as well as more long-term solutions for wide-spread access through TV white spaces to provide high-speed local Internet connectivity in rural communities.

In Chapter 6 we describe a solution that addresses the need for low-cost local cellular networks. To this end we design Kwiizya, an open-source software and hardware system that leverages a generic IP backbone to provide voice connectivity and text messaging. Kwiizya makes use of open-source implementation of GSM, which allows users to associate with their existing phones and SIM cards without any modifications. The GSM communication between users and the network is translated to VoIP by the base station, which allows Kwiizya to use generic IP backbone as opposed

to expensive commercial-grade links. For authentication and switching services, we integrate several free open-source components that handle delivery of text and calls in the networks. Where outbound connectivity is available, Kwiizya can support calls and text messages to users outside of the village network. Finally, we enhance the existing open-source software with capability to support text message-based applications through instant message-to-text message functionality.

Chapter 7 presents VillageLink, our solution to wide-spread wireless broadband in rural areas. Our focus is on TV white spaces, which with their favorable propagation properties, are a great candidate for wide-spread broadband network coverage. To this end we design VillageLink, a combined PHY/MAC mechanism for efficient channel allocation. VillageLink is designed to operate in a wide frequency band (50MHz to 800MHz) and makes use of existing TV antennas to connect end users. This design poses some unique research challenges in channel allocation due to wide-band operation and diverse antenna effects. In particular, the selection of transmission frequency can impact the existence of a link. To address this problem we make use of a two-step lightweight channel probing technique to understand the channel conditions at each communicating node. This creates a large solution space that calls for efficient methods to find the optimal channel allocation scheme. We design one such scheme that guarantees real-time convergence and identifies the optimal channel allocation to minimize interference while maximizing throughput.

A critical component of Dynamic Spectrum Access solutions such as VillageLink is their ability to dynamically detect and utilize existing spectrum resources. A fundamental research problem in this context is *how to recognize ongoing transmissions from raw spectrum measurements*. Traditionally, solving this problem has been a challenge due to the noisy nature of radio signals, which in turn makes it hard to detect ongoing transmissions. We address these challenges in Chapter 8, which presents TxMiner. TxMiner

is a system that identifies transmitters from raw spectrum measurements without prior knowledge of transmitter signatures. TxMiner harnesses the observation that wireless signal fading follows a log-normal distribution, thus most of the signals emitted by the same transmitter will fall under a Gaussian curve in the dB scale. We develop machine learning algorithms based on this observation, and show that Gaussian Mixture Models with Belief propagation can tease apart transmitters from spectrum measurements. We evaluate our algorithms on real spectrum measurements, and show that TxMiner is able to identify various types of transmitters, with different physical layer protocols, in different parts of the spectrum.

Finally in Chapter 9 we introduce an application that leverages local cellular network connectivity for improved immunizations. Vaccination distribution and tracking, especially in developing regions, is an important problem that can benefit from recent technological improvements in attaining more effective distribution mechanisms. To this end we propose ImmuNet, a system that is specifically designed for improved distribution of vaccinations to infants and children under the age of five years. ImmuNet utilizes cellular network technology and allows rapid determination of immune status; reliable updates of vaccination records; and quick, targeted dissemination of vaccination availability in rural regions. ImmuNet is a multifaceted approach for harnessing existing, unmodified cell phones; for tracking human behavior to aid in rapid, prioritized vaccination; for correlating up-to-date vaccine status of all persons in a coverage area; and for disseminating vaccination-related information to patients through local voice and data cellular connectivity.

Chapter 6

Connectivity: Kwiizya – Local Cellular Network Services in Remote Areas

6.1 Introduction

It is clear that mobile phone usage has become one of the most prevalent means of communication worldwide. 2011 statistics from the International Telecommunication Union indicate that, worldwide, there are 85.7 mobile-cellular subscriptions per 100 inhabitants [71]. This number has been steadily increasing over the last 10 years, with growth over the last five years primarily driven by subscriptions in the developing world.

While the statistics are encouraging, what they mask is the huge differential in cost and service availability and quality between the developed and the developing world. For instance, while residents of developed countries spend on average 2% of their monthly income on cellular service, the cost in developing countries is closer to 12% [71]. In addition, the ability of residents in developing regions to access cell-phone technology is reduced by limited cellular deployments; while many residents own phones and buy either subscriptions or pre-paid plans, coverage may be spotty or

non-existent within residential areas of low population densities*, as is characteristic of the majority of the world's rural developing regions. Further, while 4G is rapidly becoming available throughout the developed world, in developing regions coverage is often limited to 2G, or 2G + EDGE at best. This discrepancy hints that despite technological advancements, there will always exist an economic challenge in deployment of high-end cellular technology in sparse rural areas. Thus, there will be recurring need for lower cost solutions to serve those areas.

Despite these limitations and high costs, mobile phones are critical for providing communication in developing regions due to limited or non-existent telecommunications infrastructure and poor roads, making quick distance travel difficult. In many developing communities, cultures are oral – communication is based on oral, as opposed to written, forms. Storytelling in such communities is vital to forming world views, maintaining trans-generational knowledge and teaching practical skills. Further, oral communication facilitates practical information exchange, such as crop prices, health care availability, and numerous others. In our own fieldwork in Macha, Zambia, we have witnessed that cellphone availability combined with Internet access can enable rapid information dissemination, whereby person(s) with Internet access serve as information gateways to other, disconnected interested parties.

Our work is in partnership with the LinkNet organization in Macha, Zambia. Our team has visited Macha over the past three years, most recently spending three weeks in the Macha community during June/July 2012. During these visits, through both informal conversations and structured interviews, we have learned first hand of the deep desire of local residents for better cellular access. While every resident we encountered owns a cell phone, few have coverage at their homes. In fact, many residents own multiple phones to leverage any available coverage they come across as they travel. Further,

*The World Bank 2012 report on "Maximizing Mobile" notes that mobile network operators find it commercially infeasible to operate in rural areas.

the cost of cellular access is prohibitive. Residents rely on pre-paid access cards, which they can afford at unpredictable times due to unreliable income streams. Even the cost of frequent SMS is out of reach for many residents. Yet, the residents have enough experience with this technology to appreciate its value; 100% of our interviewees expressed a desire for better, more affordable cellular access. As a result of these costs, call durations are short, and conversations are quick and to the point. Many users simply carry phones, waiting for someone to call them, as incoming calls are free in the current cost structure.

Studies of locality of interest [76, 112] indicate that communication through technology largely appears between individuals in close physical proximity. Onnela et al. analyzed 72 million calls and 17 million text messages in Europe and found that probability of communication decreases by 5 orders of magnitude when distance between communicating parties increases from 1km to 1000km [112]. Our own prior work has demonstrated the propensity of rural users to communicate with those physically nearby: Facebook Instant Message analysis in Macha showed that, while only 35% of observed users were in the village, 54% of instant messages were between local users [76]. These statistics emphasize the need for solutions that support local voice communication in remote areas.

To address the lack of reliable and affordable cellular coverage in remote communities, we have endeavored to create a cellular access system that facilitates free local calls and SMS within a community. Our system, Kwiizya[†], leverages open-source software solutions, such as OpenBTS[‡] and FreeSwitch[§], to provide local voice and text messaging services. In our recent trip to Macha, we deployed two Kwiizya instantiations and gave pre-registered SIM cards to project partners. As a result, we have had the

[†]Kwiizya means "to chat" in Tonga, the native language in Zambia's Southern province.
[‡]`http://wush.net/trac/rangepublic`
[§]`http://freeswitch.org/`

opportunity to assess usage in a fairly unregulated, in-situ scenario. In an earlier work we outlined VillageCell, an idea for providing local calls within a community [20]. In contrast, our current solution utilizes different open-source software components, offers SMS and IM-to-SMS functionality, has been fully deployed and tested in-situ in Macha, Zambia, and opened to a set of test users in Macha.

This paper makes several contributions:

- design and integration of a *low-cost* cellular network alternative that is well suited for local communications in remote rural communities,
- evaluation of a fully operational instance of our system deployed in the field,
- modification of OpenBTS to support Instant Message to SMS functionality for SMS-based applications,
- evaluation of our platform for SMS-based applications, and
- discussion of infrastructure and sustainability-related aspects of alternative cellular network deployments.

We begin by describing the current state of rural telephony in our partner communities, including results from our detailed interviews with Macha residents. We then describe the Kwiizya architecture, including voice and SMS functionality. In section 6.3 we describe our deployment within Macha, including some of the deployment challenges we faced in this remote community. We describe results from controlled field tests, as well as results from public usage, of Kwiizya in section 8.5. Based on our performance analysis and our experience in Macha and in maintaining the system remotely, we offer our thoughts for improvement and some general guidelines for work in remote regions in section 8.7.

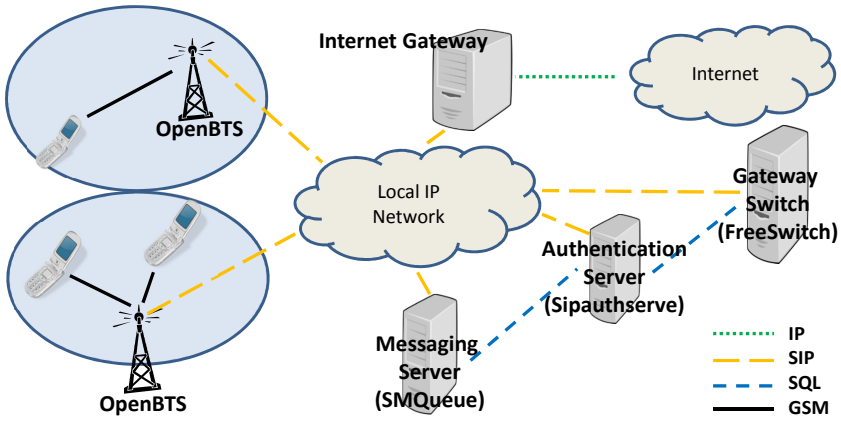


Figure 6.1: Kwiizya architecture.

6.2 Kwiizya

Kwiizya is a low-cost cellular system that leverages existing, unmodified mobile phones and SIM cards to provide voice and text messaging for free within the network. We envision Kwiizya as an underlying infrastructure for low bandwidth applications that make use of short text messages and therefore enhance Kwiizya with functionality that enables such applications. In the following section we describe the voice and text messaging operation of Kwiizya, present the integration of Kwiizya's core components, and discuss the extensions we have implemented to facilitate SMS applications.

6.2.1 Architecture

Depicted in Figure 6.1, Kwiizya utilizes free open-source software to provide voice and text message services. The base stations run OpenBTS, which implements the GSM stack and communicates with the associated cellphones using the standard Um radio interface for commodity 2G and 2.5G cellphones. OpenBTS is also responsible for translating GSM messages to SIP, which allows the use of low cost generic IP back-

bone infrastructure as opposed to an expensive commercial-grade GSM backbone. SIP translation enables the use of free VoIP server software to serve as a Mobile Switching Center for routing calls.

To route calls within and outside Kwiizya, we use FreeSwitch. FreeSwitch connects to OpenBTS via SIP and RTP and routes calls both in intra- and inter-BTS local scenarios. It has the capability to route calls outside of the network to commercial cellular, fixed line and VoIP networks using SIP and SS7. By the means of custom python scripts, FreeSwitch allows extension of the basic routing functionality to facilitate cellphone based applications. We describe in more detail the specific extensions we have implemented in section 6.2.2.

Kwiizya utilizes Sipauthserve and SMQueue to handle user authentication and text messaging, respectively. SMQueue is the SIP-based equivalent of an SMSC (Short Text Messaging Central) in a commercial-grade system. As such it interfaces with OpenBTS and makes use of commodity IP networks to transmit SMS (Short Message System). At the same time it can interface with commercial SMSCs using SS7 and SMPP. SMQueue implements a store and forward SMS queue functionality that allows messages to be delivered in a delay tolerant fashion. The latter is of great importance for areas with intermittent cellphone access and electric power availability as users are often either out of range or have their cellphone powered off. To handle user authentication and mobility, Kwiizya leverages Sipauthserve – a database server with an interface to process SIP REGISTER messages to track mobility. Both SMQueue and Sipauthserve are queried by other network elements (e.g. FreeSwitch and OpenBTS) through SQL.

6.2.2 Support for SMS-based applications

Because cellphone communication is an increasingly popular method for residents of remote areas to stay connected, cellphones are widely used for applications other

than plain voice and text messaging. Low data rate applications that leverage 160-symbol text messages have been utilized in education [82,117,126], web browsing [40], agriculture [45,115,116] and health care [21,46,47,90,114].

In the design of Kwiizya we are inherently interested in enabling functionality that will support such applications. Beyond simple SMS, we need functionality that supports SMS broadcast and multicast, and in particular does so through an Instant Message (IM) interface. Our IM-SMS interface enables fast typing and rapid outreach to a large set of subscribers. This interface has a wide variety of uses. For instance, a health worker could use it to notify subscribers of a change in health post hours of operation, or of the availability of particular vaccinations during a health scare. Our solution exposes an API that allows development of applications that can leverage the text broadcasting functionality to automate message generation, obviating the need of an actual person to send IM messages. This capability facilitates a variety of automated services, such as automatic weather alerts, dissemination of crops prices, and health care availability updates. Our design unifies this IM extension with SMS to enable rapid distribution of instant messages from any packet data network. This unified mode of messaging is asynchronous by design to minimize resource waste and guarantees only best effort delivery. We have identified four key design considerations for such a unified mode of messaging to work seamlessly with Kwiizya in a rural setting:

- A small memory footprint of the IM client on the host device; the client is designed to work on unmodified commodity hardware available in rural areas.
- Usability across various operating systems to target a heterogeneous set of devices.
- Capability of exchanging packet data with Kwiizya remotely over the Internet.

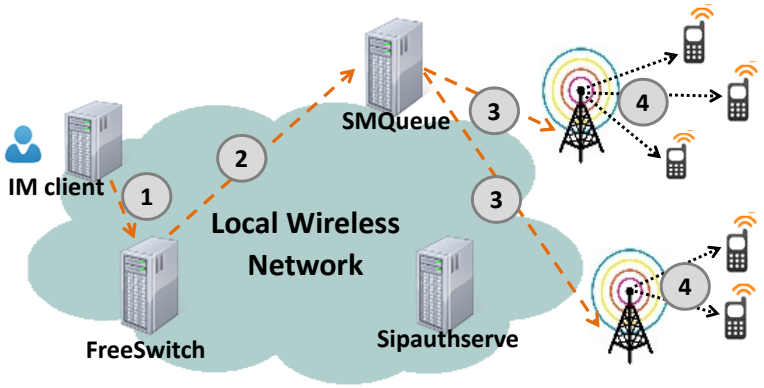


Figure 6.2: IM-SMS system architecture.

- Ability to leverage the existing SIP switching/routing capability available in FreeSwitch to enable easy and accurate routing of IM messages to Kwiizya users in the form of SMS and vice-versa.

We implement our IM client using the open-source multimedia communication library PJSIP[¶]. PJSIP implements the SIP protocol stack and supports all three NAT traversal functionalities, i.e. STUN, TURN and ICE. This facilitates routing of SIP traffic from various networks, including private IP networks such as home, office and enterprise LANs. It also exposes all functionality in suitable APIs for a wide variety of systems, including desktop and smart phones, and meets the key design goals identified above.

Figure 6.2 presents a typical usage scenario of our IM client. The client needs to be installed on a machine that can access the Kwiizya core network. The IM client then communicates directly with FreeSwitch when sending text messages to users associated with Kwiizya. FreeSwitch handles the delivery of a text message to the end user.

Provisioning a new user for Instant Messaging: Any new user of the IM service is manually provisioned with a valid SIP address of record [130] in the Kwiizya

[¶]http://www.pjsip.org/

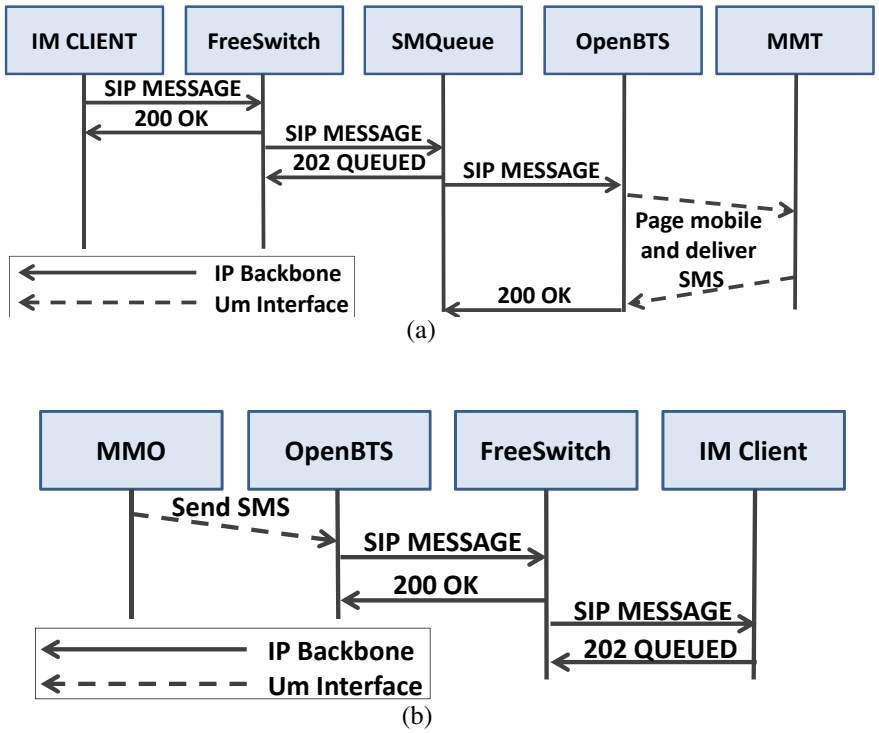
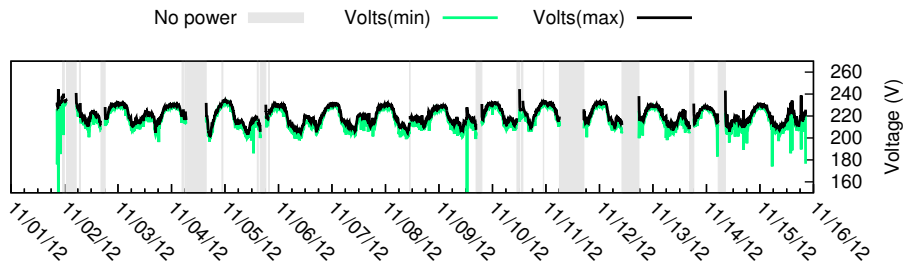


Figure 6.3: Message exchange between Kwiizya entities to send a SMS (a) from the IM client to a Kwiizya subscriber and (b) from a Kwiizya subscriber to the IM client.

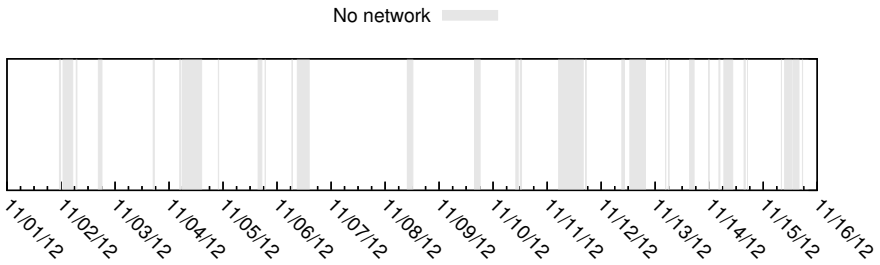
Subscriber Registry (Sipauthserve on Figure 6.2). An IM user in Kwiizya is identified by their SIP address of record (SIP user identity) and unique phone number. The SIP identity is used by internal subsystems in Kwiizya, i.e. FreeSwitch, SMQueue and OpenBTS, to validate and route any SIP traffic related to a particular IM user. The IM client's phone number is exposed to the other cellphone users of Kwiizya, so that the sender of a message can be identified and the recipient can respond if needed. To reach the IM client, Kwiizya also stores the IP address of the host where the IM client is installed. Currently, this necessitates an IM user to utilize the IM service only from a pre-provisioned host address. This mechanism, however, can later be extended to a dynamic address binding scheme.

Routing an IM message: As illustrated in Figure 6.3(a), upon sending a message to a Kwiizya user (Mobile Message Terminator – MMT), the IM client generates a standard SIP MESSAGE request [35], that contains the message in plain text format as well as the SIP address of record of the recipient. This SIP MESSAGE is then sent to FreeSwitch, which forwards it to SMQueue after validating the identity of the sender. SMQueue stores the message in its store-and-forward queue and attempts best effort delivery to the Kwiizya base station where the recipient is registered. Because the entire SMS functionality in Kwiizya is asynchronous, a real time session is not required for transmission of the IM message to its recipient. This design makes Kwiizya messaging particularly suitable for rural environments, where a user is not likely to be associated with the network at the time the message is sent. If the user is not associated, SMQueue stores the message and regularly attempts delivery. This design avoids waste of resources typically associated with synchronous/session based messaging.

Our design enables a cellphone user (Mobile Message Originator – MMO) to send a message to a provisioned IM client using the SMS interface available in a legacy GSM phone. After receiving a SMS destined to an IM user, a Kwiizya base station constructs a new SIP MESSAGE request with the SMS content attached as a SIP message body in the SIP request and forwards it to FreeSwitch (Figure 6.3(b)). After validating that both the sender and the receiver are provisioned in Kwiizya, FreeSwitch sends a new SIP MESSAGE request to the IM client's host using the IP address it stored earlier in the Kwiizya Subscriber Registry. This exchange of messages between a Kwiizya user and an IM user also occurs asynchronously between all the participating entities.



(a)



(b)

Figure 6.4: Power and network quality in Macha.

6.3 Macha deployment

During our trip to Macha in June/July 2012, we deployed an instance of Kwiizya. We chose Macha for our first field deployment, even though commercial cellphone coverage is available in the village center, because Macha is fairly well connected in terms of Internet access. This allows us to have remote access to our system for administration and performance evaluation purposes. In the following sections, we detail challenges we faced, many of which are not found in developed world deployments. We then describe how these challenges influenced our system deployment.

6.3.1 Design challenges

Power. Our Kwiizya base stations require 12VDC and draw a maximum of 3A, which can be supplied by a 12V power transformer powered by the main grid. Our

deployment locations are connected to the national power grid; however, the quality of power varies drastically over time as shown in Figure 6.4(a). The figure represents a sample two week period. The x-axis of Figure 6.4(a) plots time in days and the y-axis plots voltage. The gray areas on the plot indicate periods of power outage. We measure power quality in Macha sampling every second; each plotted voltage value corresponds to five minutes and presents the minimum (green line) and maximum (black line) measured voltage within these five minutes. The supplied voltage in Zambia should be 220V; however, as Figure 6.4(a) indicates, the voltage varies from 150V to 240V with frequent power failures. Furthermore, there are long periods, for example November 15 and 16, of power brownouts in which electricity is available, but the voltage is continuously low. This is particularly harmful for computer equipment. Voltage tends to follow a diurnal pattern, typical for a system that is overloaded; the power quality is particularly poor during the day when the utilization of the power grid is higher. This poor power quality is harmful to equipment and makes remote access for administration and evaluation extremely challenging.

Internet access. The remote accessibility of the system is not only influenced by power availability but by network availability as well. Power is often available in the village; however, due to an outage in the upstream link, the village gateway cannot connect to the Internet. Figure 6.4(b) shows network availability over the same two week period in November. In comparison to power outages, there are many more network outages, which further reduce our ability to access Kwiizya remotely.

Logging and storage. Logging is important for system administration and troubleshooting; however, special attention should be paid when enabling logging to ensure it does not deteriorate system performance, as writing to disk can slow the system. We experienced an event where a combination of detailed logging and limited storage in the base station caused Kwiizya to malfunction.

(a) (b)

Figure 6.5: Our equipment in Macha: (a) the base station and (b) the power supply.

We adapt our system design to meet these challenges. The following section provides technical details about our field deployment.

6.3.2 Technical details

Our Macha deployment operates in the GSM-1800 band. We chose this band for two reasons. First, most residents have basic dual band (900/1800) phones, so the use of GSM-1800 maximizes the number of people who will be able to use the system with their existing phones. Second, GSM-1800 does not interfere with commercial cellular providers, which operate at GSM-900 in rural areas.

Kwiizya consists of two self-contained RangeNetworks base stations[‖] (Figure 6.5(a)), each of which covers a radius of up to 3km. Each base station unit includes a software-defined radio, a 1W power amplifier, a duplexer, and a small integrated PC with 4GB flash card that runs Ubuntu 10.04. Each unit can independently run all components of a GSM network. However, to scale the network to a wider physical area, we use one of the base station PCs as a network central server, running FreeSwitch, Sipauthserve and SMQueue; the second base station runs only OpenBTS and connects to the first one (Figure 6.6) for all other services. The connection between the two base stations

[‖] `http://rangenetworks.com/`

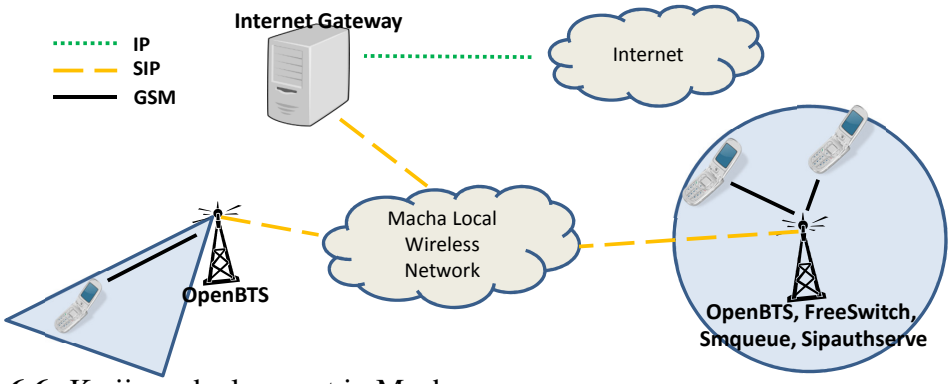


Figure 6.6: Kwiizya deployment in Macha.

is facilitated through a long distance Wi-Fi link that was readily available through the local wireless network in Macha.

To handle the problem of limited storage we attach a 1TB external hard drive to each base station. We check the internal disk utilization periodically and when it reaches 90%, we offload syslogs from the base station PC onto the external hard disk to free disk space and save logs for future reference. We use simple ping based monitoring to inform us when access to the village gateway is available from the outside, so that we can analyze network availability and know when our logs are accessible.

Determination of the most suitable power supply to resist the power fluctuations in Macha while providing stable 12V/3A to our base stations was an iterative process. We started with a complex of deep cycle car batteries and car battery chargers; however, within ten days we lost one charger due to poor power quality and two of the batteries started leaking due to overcharge. Eventually, we changed our power supply to the one shown in Figure 6.5(b), which includes a UPS, a 5A/12V monitor power supply and a wooden block with nails to keep the power cable terminals in place, so they do not short out.

(a) (b)

Figure 6.7: Kwiizya sites in Macha: (a) the water tower and (b) LITA.

The area of Macha is relatively flat with small hills, so providing coverage was a matter of installing the base stations on elevated locations. We installed one of the base stations on a 30m water tower with a 10m communication mast on top (Figure 6.7(a)). For this site we used an 11dBi omni directional antenna. Our second base station was installed close to the LinkNet IT Academy (LITA) on a 12m mast mounted on the ground (Figure 6.7(b)) and used an 11dBi 90 degree sector antenna. The distance between the two sites is 2.3km; each site is capable of providing coverage in up to 3km radius depending on terrain.

While Kwiizya supports open registration with existing cellphones and SIM cards, we opted for restricted registration during our initial field testing. We manually provisioned 20 SIM cards and distributed the cards to a small set of users comprised by our local partners and their families. This allowed us more control during the initial performance evaluation of our system, while still allowing our users unrestricted access. In the near future, we plan to deploy Kwiizya in neighboring communities with no existing cellphone coverage and enable open registration.

We placed two GSM modems in the field, which allows us to send SMS and conduct calls in the Kwiizya network in a controlled manner. We chose U-Blox quadband GSM/GPRS modems that can be powered and controlled through USB. The latter is important because it allows the modems to be powered by the server to which they are connected, thereby making them independent of an individual reliable power supply. We attached the modems to an Ubuntu server and were able to access and control them remotely using AT commands, which are a suite of specialized commands for remote control of GSM modems.

Each call in Kwiizya has two associated VoIP sessions – one for the mobile call originator (MCO) and one for the mobile call terminator (MCT). A VoIP session consists of a SIP control session and forward and reverse RTP streams that carry the actual voice traffic. The Kwiizya deployment in Macha supports four call scenarios depending on the locality of the MCO and the MCT. We describe these call scenarios in turn:

- *Water tower – Water tower.* This is a call scenario where both the MCO and the MCT are in the vicinity of the water tower base station. This setup does not utilize the backhaul wireless link; both VoIP sessions are established only through the water tower base station.
- *Water tower – LITA* and *LITA – Water tower.* In the first scenario, the MCO is associated with the water tower and the MCT with LITA; the opposite holds for the second call scenario. These two scenarios are identical in terms of resource utilization. In each of the cases one of the VoIP sessions utilizes the backhaul link and the other one is local at the water tower.
- *LITA – LITA.* In this case both the MCO and the MCT are attached at LITA. In this scenario, both the MCO and MCT VoIP sessions traverse the wireless backhaul.

6.4 Evaluation of Kwiizya

Our evaluation of Kwiizya includes two parts – (i) controlled experiments and (ii) evaluation of system usage in the wild following our Macha deployment. Our controlled experiments take place in two testbeds: one located in our lab in UCSB and one in Macha. We first evaluate Kwiizya by running a number of controlled experiments that test the system's ability to provide services and the quality of these services. The second part of our evaluation assesses the actual deployment experience.

6.4.1 Controlled experiments

We focus our controlled experiments on three aspects of the system: (i) text messaging, (ii) voice calls and (iii) instant messaging to SMS. In the process of evaluation we take into account OpenBTS' capabilities as well as limitations of the GSM modems to ensure the accuracy of our performance measurements. First of all, OpenBTS can handle at most 84 SMS messages per minute and 7 simultaneous calls, so we ensure we do not generate traffic at a higher rate so that delay and call quality are accurately measured. Furthermore, the GSM modems incur a delay of about 0.5 seconds to execute the commands associated with sending a single message. Finally, there needs to be an inter-message delay of 6 seconds to ensure the GSM modem does not become overloaded with outgoing messages. When overloaded, the modem fails to send messages to Kwiizya.

We conduct our controlled experiments in two testbeds, one located in our lab at UCSB and one located in Macha. Our lab setup includes one RangeNetworks SNAP unit, the same as the ones deployed in Macha. The unit operates as a self contained system running all four services – OpenBTS, FreeSwitch, Sipauthserve and SMQueue.

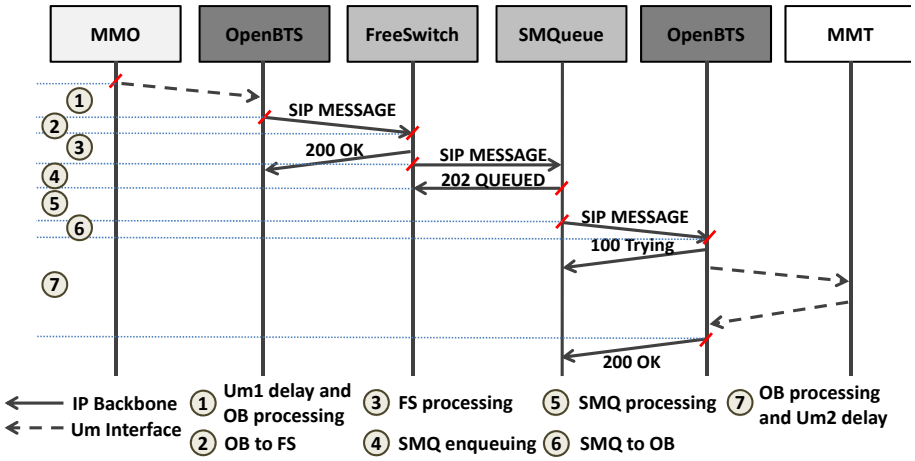


Figure 6.8: SIP messages transmission of a SMS from MMO to MMT.

Our testbed in Macha includes the base station installed at LITA and a second base station inside the IT room. The indoor base station runs FreeSwitch, Sipauthserve and SMQueue. LITA runs only OpenBTS and connects to the indoor base station for the other services. In this test setup we utilize the readily available network infrastructure in Macha to connect the two base stations. As a result, we use an existing wired connection for the backbone link between the indoor installation and LITA. All Kwiizya components are strictly time synchronized. This is achieved using NTP, where all machines in Kwiizya synchronize to a local server to maximize accuracy.

6.4.1.1 SMS and Voice Calls

To evaluate SMS and voice call quality through controlled experiments we use our testbed in Macha. We deploy two U-blox quadband GSM/GPRS modems in Macha and associate them with the LITA base station. Both modems are connected to a server that can be accessed from the Internet. We then automate SMS and voice calls between the two modems.

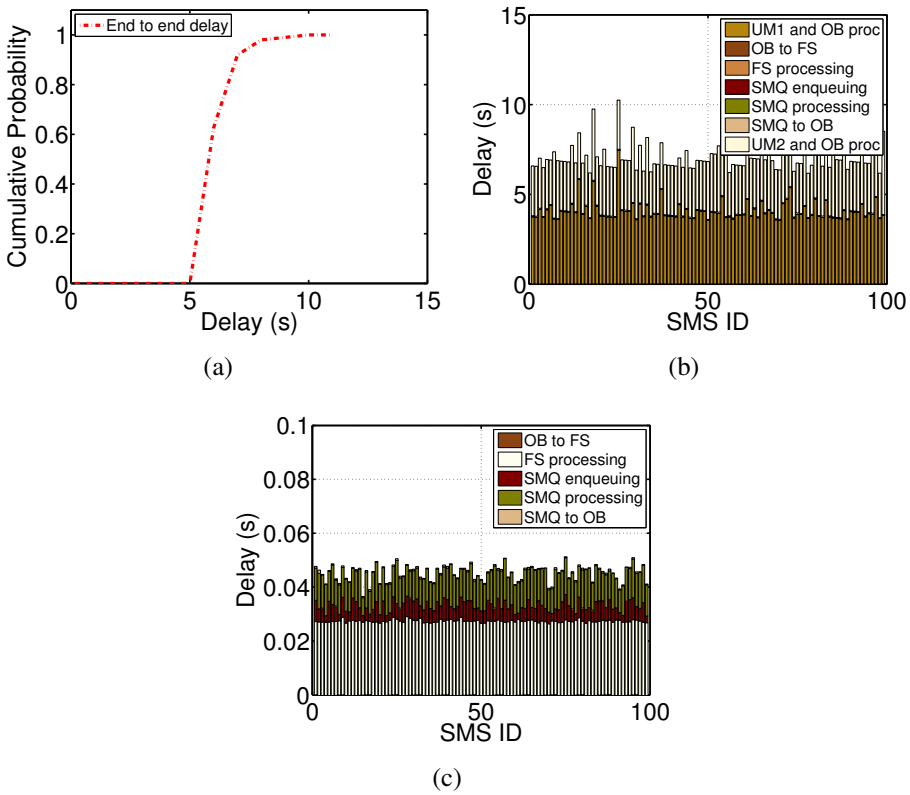


Figure 6.9: Evaluation of 100 SMS transmissions to associated users. (a) CDF of end to end delay; (b) breakdown of delay components per message; and (c) zoom of non-Um delay components.

We evaluate the end-to-end delay for delivery of a single message and assess the delay components incurred by each element of Kwiizya. Figure 6.8 gives an overview of the SIP message exchange associated with delivery of a single SMS from a Mobile Message Originator (MMO) to a Mobile Message Terminator (MMT). If two components are coded with the same color, they run on the same physical machine. The figure also presents our approach to calculating the delay components. We calculate the end to end delay as a sum of each extracted delay component.

In our evaluation of SMS delivery we focus on two scenarios. First, we evaluate delay for delivery of messages to users associated with the network. We then evaluate the impact of SMQueue load on the end-to-end delay by increasing the number of transmitted messages to users that are provisioned but not associated with the network. The latter increases the load of pending messages in SMQueue. We then measure the time it takes for a message to be received by a registered user.

To evaluate the delay for message delivery to a user associated with Kwiizya, we sent 100 consecutive messages, with 6 seconds inter-message delay. 99 of these messages were received by Kwiizya; one message failed to depart from the sending GSM modem. Of the 99 messages that entered Kwiizya, all 99 were successfully delivered to the receiving modem. Figure 6.9 presents our results. In Figure 6.9(a) we plot a CDF of the end to end delay. As the graph shows, 99% of the messages were delivered to their destination in $6 - 9$ seconds; the maximum observed delay is 11 seconds. The average delay over all 99 messages is 7.06 seconds with standard deviation of only 0.7 seconds. Figure 6.9(b) presents a breakdown of delay incurred by the separate Kwiizya components for each text message. The major contributors to the end-to-end delay are the radio interfaces from the MMO to OpenBTS (Um1) and from OpenBTS to the MMT (Um2). Figure 6.9(c) shows a zoom of the delays of the non-Um components of Kwiizya. The contribution of the non-Um components together is at most 50ms with

average over all messages of 45.4 ms. This performance is persistent over different messages, as the standard deviation for non-Um delay is only 2.9 ms.

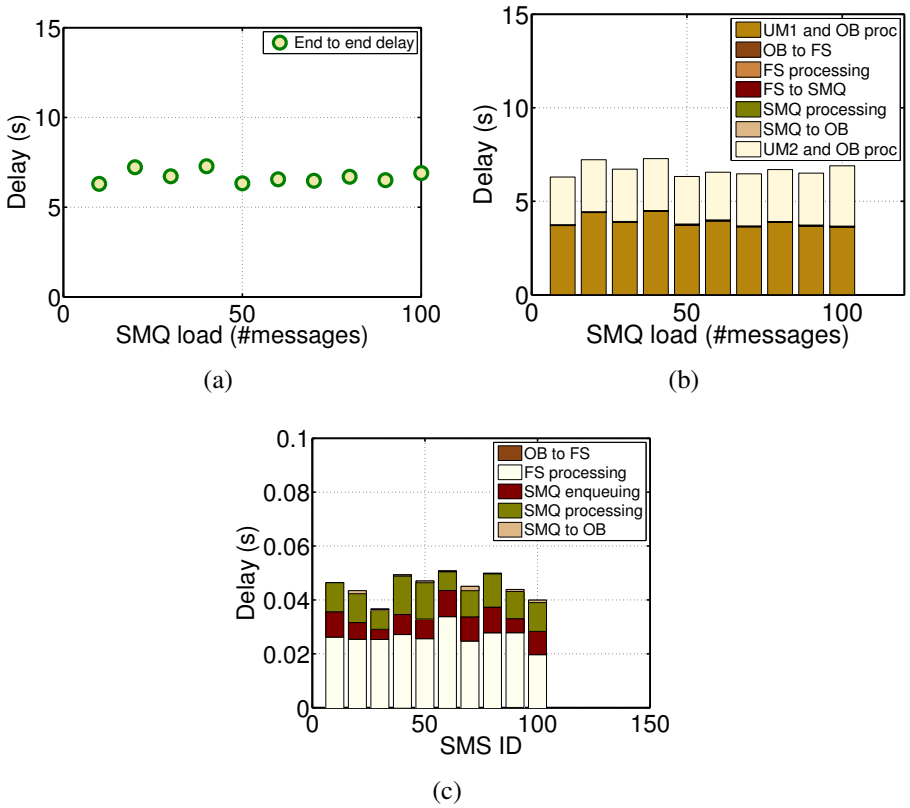


Figure 6.10: Impact of increasing SMQueue load on (a) the end to end delay; (b) the delay components; and (c) non-Um delay components.

Where users do not have continuous access to Kwiizya, the network stores text messages and aggressively attempts to deliver them every second. This incurs overhead in SMQueue in storing messages and reattempting delivery. We evaluate the impact of this overhead on the end to end delivery delay of messages to registered users. We increase the SMQueue load from 10 to 100 messages in steps of 10. At each step we send one message to a registered user and measure the delivery delay for this message. As Figure 6.10(a) shows, the end to end delay does not increase with the number of enqueued messages; thus it is not dependent on the SMQueue load for as much as

100 enqueued messages. This observation is further confirmed by Figures 6.10(b) and 6.10(c), which show delay components for each of the 10 delivered messages. Once again, Um1 and Um2 are the largest contributors to the end-to-end delay. Furthermore, each of the ten messages is delivered in less than 8 seconds, which is within the average delay range encountered on Figure 6.9(a).

Finally, in our test setup in Macha, we measure the call establishment time. As depicted in Figure 6.11, this is the time from when the first SIP INVITE message is sent by the MCO base station (informally from when the calling party hits "dial") to when the MCT starts ringing (or when the calling party starts hearing the "ring" tone). Again components with the same color run on the same physical machine. We run an experiment with 15 consecutive calls. Figure 6.12 presents our results; call ID is plotted on the x-axis and call setup delay in seconds on the y-axis. The figure indicates that with one exception, it typically takes between 2 and 4 seconds for a call to initiate. We examined the delay components of the outlier call 14; the 15 second setup time is caused either by delay in the Um radio interface between OpenBTS and the GSM modem or by a glitch in the GSM modem itself. We leave the evaluation of call quality to the calls placed by actual users; our findings are described in section 6.4.2.

6.4.1.2 Instant Messaging to SMS

As detailed in section 6.2, we extend Kwiizya's basic cellular network functionality to provide a platform for development of applications that utilize text messaging. This platform delivers an interface between an instant messenger and the cellular network and enables burst message transmission to users registered with the network.

We evaluate our Instant Messaging to SMS (IM-SMS) functionality in our lab setup at UCSB. We perform this evaluation at UCSB because of the need for multiple test receivers, for which we did not have the capacity in Macha. We installed our IM client

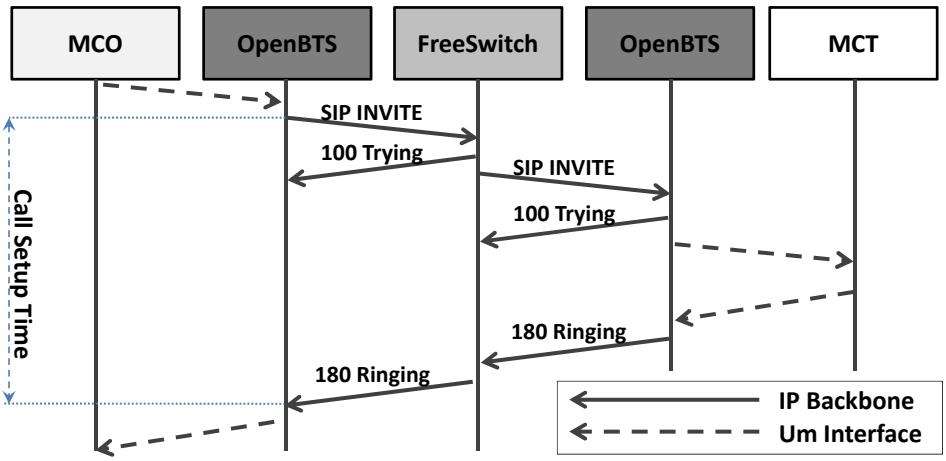


Figure 6.11: Call setup SIP message transaction.

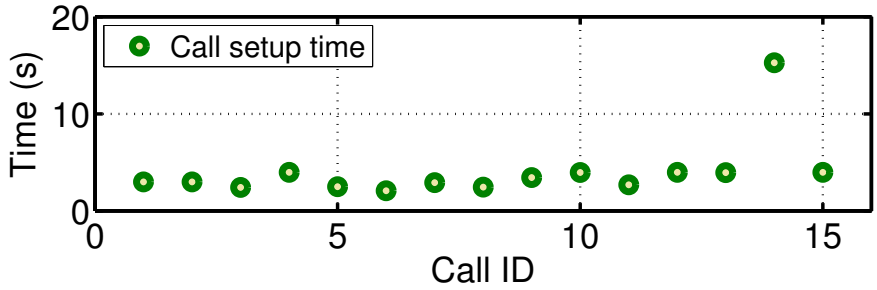


Figure 6.12: Call setup time.

on a server, running Ubuntu 12.04. We associate three HTC G1 phones with the test network to act as receivers during our experiments.

We conduct two sets of experiments to evaluate the end to end delay for message delivery from the IM client to users in Kwiizya. We also calculate the delay imposed by each Kwiizya component. Figure 6.13 presents the SIP message exchange for delivery of a SMS and breaks down each delay component. Again components that run on the same physical machine are coded with the same color.

Our first experiment evaluates the time for message delivery when the IM client sends bursts of messages to *associated users*. In particular, we are interested in the implications of the burst size on the end to end delay. We send bursts of messages of in-

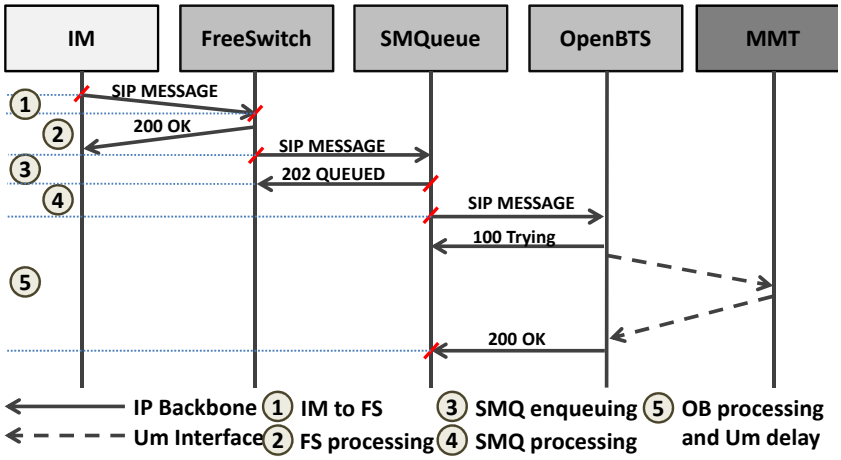


Figure 6.13: SIP messages transmission of a SMS from IM to a Kwiizya user.

creasing size from the IM client to three users associated with Kwiizya with inter-burst time of 60 seconds. At each iteration 1/3 of the burst is sent to each of the receivers. It is important to note that each phone can only receive one message at a time. Thus, if five messages are sent simultaneously to the same phone, the first message should arrive in X seconds, the second in $2X$ seconds and so on. In our IM-SMS tests we used three phones to receive bursts ranging in size from 15 to 75 messages. Every message sent in this experiment was received at its destination.

Figure 6.14 plots burst size on the x-axis and end to end delay on the y-axis. Each point on the plot corresponds to delay for a single message and each vertical line indicates the burst size. In each segment of the graph there are $burstSize/3$ well established groups, each of which contains three plotted values, one for each receiver. On the plot these three values appear overlapped, which indicates that message delivery to all three phones is approximately the same. This separation in groups is due to the increasing delay for delivery of a burst of messages to the same user. Messages across different burst sizes take the same amount of delivery time, which demonstrates that Kwiizya is capable of handling bursts of variable size without incurring delay in mes-

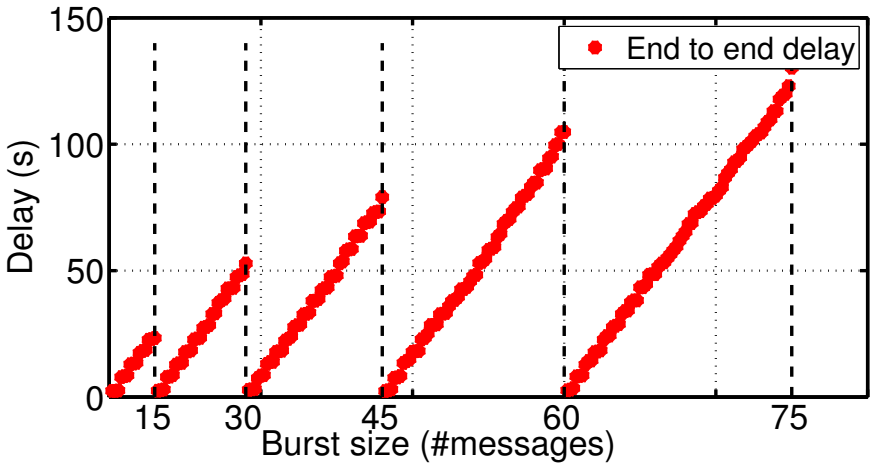
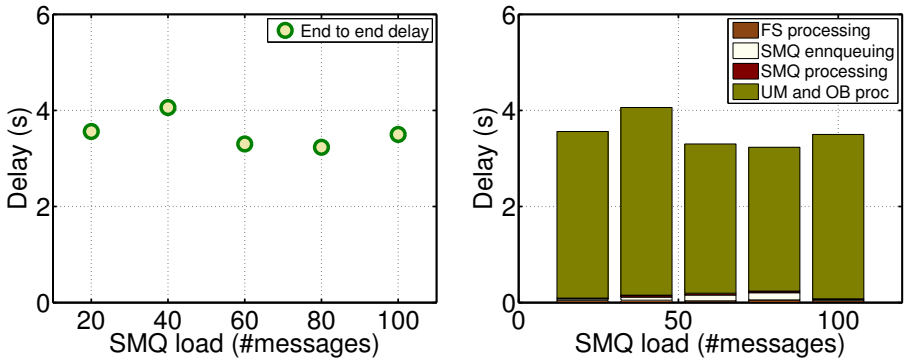


Figure 6.14: IM-SMS end to end delay when sending to registered users.

sage delivery. In a production network, it is unlikely that a single phone would need to receive large bursts. If the number of recipient phones corresponds to the size of message burst sent by the IM client, we will observe a constant delay across burst sizes, which will be close to the minimum measured delay – between 2 and 4 seconds.

In the second experiment we evaluate the impact of SMQueue load on the end to end delay during message bursts. To do so, we send bursts of messages to a mix of associated and non-registered users. To generate load in SMQueue, we transmit an increasing number of messages to a non-registered user, starting at 20 messages, and add 20 messages at each of five iterations. At each SMQueue load level we also send a burst of three messages – one to each of the three phones associated with Kwiizya. We measure the end to end delay and delay components for delivery of these three messages. Figure 6.15 presents our results. In 6.15(a) we plot the average end to end delay for the three deliverable messages sent at each iteration, as a function of SMQueue load. The delay does not increase with the increase of SMQueue load up to a hundred messages, which demonstrates Kwiizya's capability to deliver bursts of messages without negative impact from the number of enqueued messages in SMQueue. In Figure 6.15(b) we



(a) (b)

Figure 6.15: Impact of SMQueue load on message delivery delay from IM to a Kwiizya user: (a) end to end delay and (b) breakdown of delay components.

plot the delay components averaged over the three deliverable messages as a function of the SMQueue load. As we can see, the delay caused by the Um interface is still the largest delay component, while the one introduced by SMQueue is negligible.

6.4.2 Kwiizya field usage

We now present results from in-situ usage of our Kwiizya deployment in Macha from a period of two weeks in July 2012. We begin by describing our traffic collection system. We then evaluate call quality and compare with results from our controlled experiments. We note that the backhaul Wi-Fi link in our deployment in Macha carries live Internet traffic and thus there may be variability in results due to competing traffic. The impact of this background traffic on call quality was evaluated in our previous work [20].

6.4.2.1 Kwiizya monitoring

As noted in section 6.3, our deployment includes two base stations – water tower (*WT*), which runs OpenBTS, FreeSwitch, SMQueue and Sipauthserve; and LITA (*LT*),

which runs only OpenBTS and connects to *WT* to use other services. To capture all system traffic, we installed three monitoring points and run tcpdump at each point to capture SIP and RTP traffic:

- *lo@WT* – the loopback interface at the water tower base station that captures all internal communication in *WT* between FreeSwitch, SMQueue and Sipauthserve. This monitoring point allows evaluation of calls where either or both the MCO and the MCT are in the vicinity of *WT*.
- *eth@WT* – the Ethernet interface at the water tower base station that connects to the Wi-Fi link from the water tower to LITA and captures all communication between *WT* and *LT*. This includes all SIP and RTP traffic related to calls where either one or both communicating parties are associated with LITA.
- *eth@LT* – the Ethernet interface at the LITA base station that connects to the Wi-Fi link from the water tower to LITA. This monitoring point captures the same traffic as *eth@WT*, however with different timing of packets. This allows us to assess timing related aspects of the system.

6.4.2.2 Voice call quality

Delay, Jitter and Packet Loss. Delay, jitter and packet loss are three characteristics of a VoIP session that are critical to voice quality. ITU recommendation G.114 mandates that tolerable one way delay is up to 150 ms [11]. The theoretical delay minimum that a system can provide is dependent on the used codec. Kwiizya uses GSM 6.10, for which the minimum delay is 20 ms. We evaluate these characteristics in our system by analyzing 52 VoIP sessions from Kwiizya users in Macha. Figure 6.16 presents our results. For each VoIP session, Figures 6.16(a) and 6.16(b) plot the average and standard deviation of delay and jitter. A single point shows the average delay or jitter and standard deviation across all packets within that VoIP session. Both

delay and jitter are well below the tolerated thresholds for VoIP. At the same time, the average delay is close to 20 ms over all VoIP sessions. Furthermore, delay and jitter do not vary much over a single VoIP session as indicated by the standard deviation bars, which shows that Kwiizya has stable performance throughout a call. Finally, Figure 6.16(c) plots packet loss over the 52 VoIP sessions. Only three of all sessions suffered non-zero packet loss; however, these three were all less than 0.5%, which is within the limit for satisfactory call quality.

Mean Opinion Score. Mean Opinion Score (MOS) is a metric that describes the call quality as perceived by the communicating parties. The maximum MOS that a system can provide is dependent on the voice codec. We evaluate the MOS achieved by Kwiizya by utilizing the E-model [10]. This model takes into account the codec type as well as experienced packet loss during a call and outputs the MOS. According to the E-model, MOS is calculated as follows:

$$MOS = 1 + R.0,035 + R.(100 - R).(R - 60)).7.10^{-6} \tag{6.1}$$

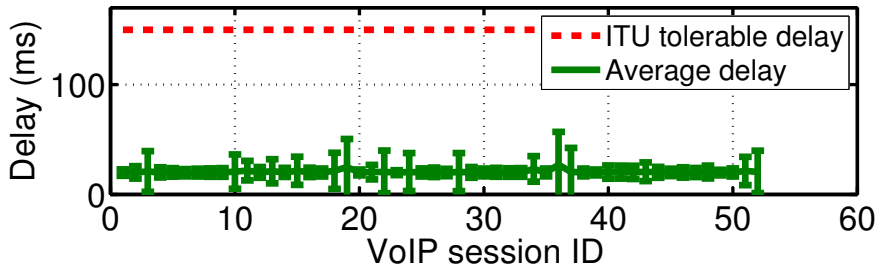
where R is the *rating factor*, calculated as a function of the *effective equipment impairment factor* Ie_{eff}:

$$R = 93.2 - Ie_{\text{eff}} \tag{6.2}$$

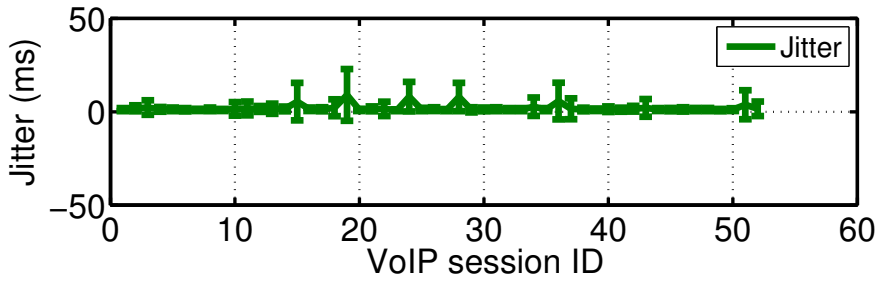
The *effective equipment impairment factor* is packet loss dependent and can be found as follows:

$$Ie_{\text{eff}} = Ie + (95 - Ie).\frac{ppl}{ppl + bpl} \tag{6.3}$$

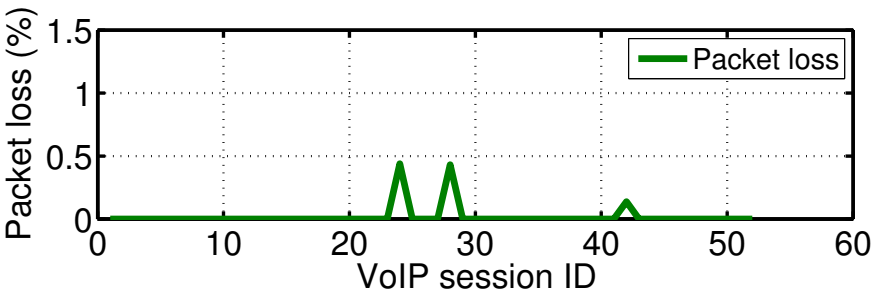
where Ie is the *equipment impairment factor* taken for zero packet loss; bpl is the *packet loss robustness factor* and ppl is the *packet loss probability*. Ie and bpl are



(a)



(b)



(c)

Figure 6.16: Field call (a) average delay; (b) average jitter and (c) packet loss.

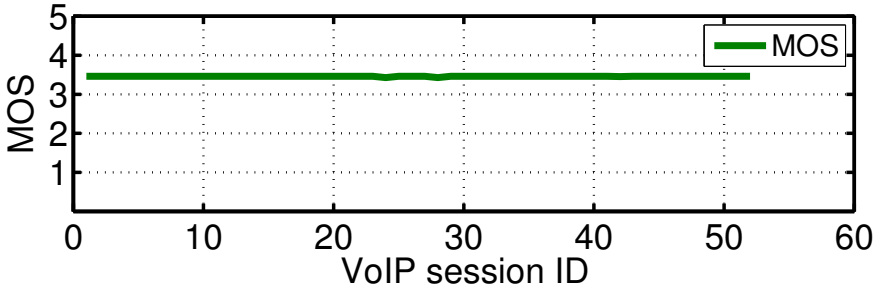


Figure 6.17: Mean Opinion Score.

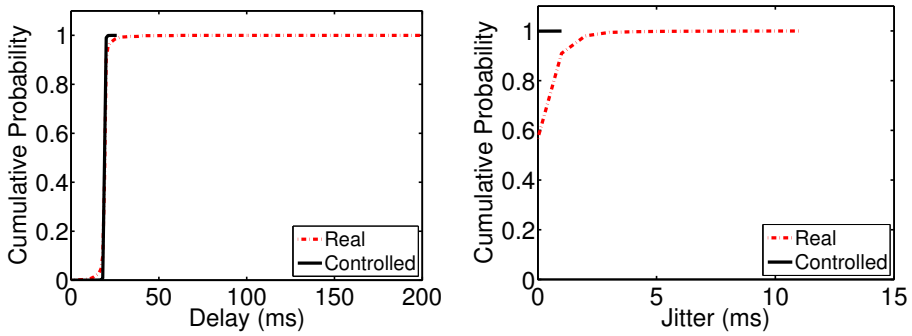
given for different codecs in a recommendation by ITU [6]. For a single VoIP session consisting of two RTP streams (one forward and one reverse), the ppl value is calculated as follows:

$$\begin{aligned} totLoss &= loss_{fwd} + loss_{rvs} \\ totSent &= sent_{fwd} + sent_{rvs} \\ ppl &= \frac{totLoss}{totSent} \end{aligned} \tag{6.4}$$

FreeSwitch uses the GSM FR (6.10) codec, for which the Ie and bpl values are 26 and 43, respectively [6]. Given these, we can calculate that the maximum expected MOS provided by GSM FR (6.10) is **3.46** (effectively this is the MOS for $ppl = 0$).

Figure 6.17 presents our results for MOS for the 52 VoIP sessions. On the x-axis we have session ID, while on the y-axis we have MOS. The minimum measured MOS is 3.43 and the maximum (received by 49 of the 52 sessions) is 3.46 – equal to the maximum provided by the system.

Controlled vs real call. We compare the performance of a call conducted in our Macha testbed and one from a user in the network. Both calls are characterized by the same packet size and inter-packet timing of RTP packets in the media streams. There are two major differences between the Macha testbed and the actual deployment. First, the testbed uses a wired backhaul, whereas the actual deployment uses a long distance



(a) (b)

Figure 6.18: Controlled vs. actual user call: CDF of per packet (a) delay and (b) jitter.

Wi-Fi link for inter-base station communication. The second difference is the distance between the users and the base stations to which they are connected; while in the testbed our GSM modems are only about 40m away from the base station, in the actual deployment phones can be anywhere in the range of a base station. Probabilistically, they are further than 40m since the transmission range is approximately 3km.

We present our comparison results in Figure 6.18. 6.18(a) and 6.18(b) plot CDFs of per-packet delay and jitter, respectively. The black solid line in both graphs presents results from our controlled call while the red dotted line presents results from a representative call from the actual deployment. We can clearly see the impact of the wireless link and the longer radio link to users' phones on the per-packet delay and jitter. While in our controlled call we experience a maximum delay and jitter of 26ms and 1ms, the actual call experiences slightly higher values. Nevertheless, more than 99% of the packets in the actual call experienced delay less than 30ms and jitter less than 3ms, which indicates that utilizing a long distance wireless backhaul link and realistic distance from users to base stations, Kwiizya can provide for delay and jitter well below the thresholds that threaten performance.

6.4.2.3 Text messaging

During our two week evaluation period, we observed only three text messages exchanged between Kwiizya users. From our conversations and interviews with cellphone users in Macha and the surrounding villages, we learned that residents typically use SMS more frequently in production cellular systems due to the lower cost of SMS in comparison with a voice call. When voice calls are more affordable, people prefer to call instead of text. This is in line with the oral communication culture within these communities, as well as high illiteracy rates. From our initial analysis of Kwiizya traffic, we observe the tendency of users to place voice calls instead of sending SMS messages because voice calls are free. However, it remains to be seen whether this observation would continue to hold once Kwiizya is opened to the general community.

6.5 Related Work

The impact of cellphone technology on residents of developing countries has been widely studied [17,45]. Examples from the literature show that cellphones have changed the way people learn and exchange knowledge, get access to health care and handle local government activities. There is increasing effort to develop applications that use plain text messaging and voice to leverage existing feature phones. These applications can be largely divided into those that use voice and those that use text messaging. In [116] for instance, the authors deploy a social network based on voice forums for farmers to exchange knowledge about crops and crop prices. [111] proposes a method for message exchange where messages are encoded as a sequence of missed call durations. Numerous applications use text messaging as a platform. Use cases include update of remote databases through SMS [90], attendance tracking [126], web search [40] and health care [21, 46, 82, 114].

The drawback of these solutions is that they all assume existence of an underlying cellphone network. However, ITU statistics show that 48% of the rural population in sub-Saharan Africa is still disconnected; this constitutes a large fraction of a population that benefits most from the outlined applications.

A few projects discuss alternative cellular network solutions for remote areas [65, 105]. Heimerl and Brewer were the first to propose the use of a PC and a SDR for cellphone coverage [65]. The paper describes an idea for providing intermittent cellular access to under-serviced areas, where the design focus is on minimizing the energy consumption for running a rural base station. Mpala et al. study the applicability of an OpenBTS based system in rural areas [105]. The paper describes a show case deployment of an OpenBTS and Asterisk-based cellular network with a limited range. These papers provide insight on design of such systems and discussion of obstacles to implementation; what they lack is holistic system design and large-scale evaluation of feasibility from a system point of view.

6.6 Discussion and Conclusion

The introduction of software-defined radios and software that converts GSM signals to voice over IP allows design of low-cost local cellular networks, which can notably improve communication in remote communities. We leverage these technologies and deploy the first large-scale prototype of such a network, Kwiizya, in the remote community of Macha in Zambia. While Wi-Fi is available in the community, access to the Wi-Fi service and Wi-Fi enabled user equipment is expensive. As a result, the majority of the people in the community only have access to feature cellphones. Kwiizya leverages unmodified cellphones to provide voice and SMS services; this capability is essential to covering large populations.

Kwiizya can scale to larger deployments through multiple options. It can use long-distance wireless links to physically connect communities that are a few kilometers apart. In cases where Internet is available, Kwiizya can use a globally accessible Private Branch Exchange server to connect individual installations through the Internet without the need of a physical connection between these installations.

Kwiizya can also interface with global services such as GoogleVoice, Skype or commercial operators, which allows Kwiizya users to call and text with other users who are outside the local network. We explored a configuration that uses GoogleVoice for outbound calls and text messaging; however, frequent changes in the GoogleVoice API as well as the inability to hide caller ID stopped us from opening outbound services to our users in this deployment. We plan to continue development of this option in the near future.

A reliable power supply is a major fraction of both capital and operating expenses in remote cellular deployments. Many remote deployments typically operate off the power grid, which requires alternative power sources such as diesel generators or renewable alternatives. 80% of the power consumption of a commercial cellphone network is induced by the base stations [94]; this constitutes a major fraction of the operation cost. This expense is further increased in the case of remote deployments due to the need for transportation of diesel or maintenance of renewable power sources. While this paper is not concerned with thorough evaluation of power consumption, we note that Kwiizya has significant advantages in power consumption in comparison with a commercial system. While a RangeNetworks SNAP unit such as the ones we use in Macha draws a maximum of 35W[**], a typical commercial-grade cellular base station[††]

[**] `http://www.rangenetworks.com/store/snap-network`

[††] We compare with a Siemens BS-240 that supports 24 simultaneous calls and has a 40W output power amplifier.

consumes 1.3kW. Thus, Kwiizya has great potential to reduce the operating expenses associated with power supply for small-scale local deployments.

While in this paper we focus on technology, there are several economic and regulatory factors beyond the scope of our work that influence the deployment of alternative cellular solutions in under-provisioned areas. Economic factors concern availability of financial resources to provision power, Internet access and equipment (software and hardware), while regulatory factors are largely related to licensing. Our partnership with LinkNet in Macha has helped us address economic aspects of Internet access and power availability. As opposed to a commercial-grade cellular system that uses licensed and costly software, Kwiizya relies entirely on free, open-source software. This, coupled with lower-cost hardware, reduces the cost for a singe base station by five to ten times in comparison with a commercial-grade counterpart. Licensing for experimental purposes is widely unregulated in sub-Saharan Africa. In Zambia, there is a national regulatory organization that governs the distribution of spectrum resources; however, there is no policy for granting experimental licenses [105]. Furthermore, pricing of commercial licenses is oblivious to the potential return of investment. We argue that lower pricing or subsidizing commercial licenses for economically unattractive areas such as rural areas would encourage entrepreneurs to seek alternative technologies and deploy in such areas. In this context, while Kwiizya has great technological potential to provide low-cost cellular infrastructure, the full potential of the system is yet to be realized. This realization largely depends on multi-faceted economic analysis of alternative solutions for cellphone access.

Deployments in remote areas require systems independent from factors such as power quality and availability, weather and network integrity. We have identified a few aspects of Kwiizya that need to be improved to make the system even more suitable for deployment in any remote community. First, a transition to solar power would

facilitate deployment independent of the power grid and provide high quality power. A follow up evaluation would then help realize Kwiizya's potential to operate in the face of intermittent power availability. Second, it is very important to weatherproof a system that is installed in the field. For example, one of our base stations in Macha was damaged by lightning; we have learned first hand the importance of good insulation and lightning surge protectors. Lastly, to handle network integrity problems and provide continuous service, Kwiizya must be self-organizing, meaning a base station needs to sense whether access to central services is available and, if not, switch to another central server or fall back into self-contained mode.

Kwiizya is the first full scale deployment of an alternative, affordable cellular communications architecture for rural areas that lack cellular coverage. Through extensive evaluation, we demonstrated Kwiizya's capability to provide high quality services even through shared backhaul. Kwiizya reduces the capital expense by using free open-source software, SDR-based hardware and universal IP backbone. Kwiizya also reduces operation expenses by incurring much lower power consumption in comparison with a commercial cellular network. Kwiizya provides an efficient platform for SMS-based applications. Part of our ongoing work focuses on the design of such applications after careful consideration of specific needs in the community. We plan to continue to work with our partners within the Macha community to extend deployment of our solution to surrounding areas that lack cellular services, to both residential and public service areas, such as rural health posts, that are currently completely disconnected. Through longitudinal studies, we hope to provide further insights on the adoption and impact of our deployment on the local community.

6.7 Acknowledgements

The material in this chapter is based on joint work with Arghyadip Paul, David L. Johnson and Elizabeth Belding. We are also thankful to our partners from Macha: Gert-jan van Stam, Consider and Pharren Mudenda, Austin Sinzala and Acent Milimo. We would also like to thank RangeNetworks for their support, Veljko Pejovic and David Kotz for their valuable feedback, and Kate Milosavljevic for her help with the interviews.

Chapter 7

Connectivity: VillageLink – Wide-Area Wireless Coverage

7.1 Introduction

Internet connectivity is available to a small fraction of the total world population. Only 39% of the population has an opportunity to go online, while in the developing world that percentage is even lower – only 31% [12]. Moreover, users in developing regions are often limited to Internet access in a publication location, such as their work place or an Internet café. In Africa only 7% of Internet users have at-home access [12]. Public access is associated with problems such as limited hours of operation, high cost, and, of course, long distances from one's home. As a result, people are forced into a so-called "deliberate interaction" model where tasks, such as e-mailing, are planned in advance, and where any freedom in exploring the Internet and all its applications is minimized [149]. This Internet access model is not limited to developing regions; many remote communities in the developed world are currently disconnected as well [138].

The main cause of limited Internet penetration stems from the fact that more than three billion people live in rural areas. These areas are hard to connect via copper cables, fiber optic or cell phone base stations due to high deployment cost and low pop-

ulation density which renders these techniques economically infeasible. Rural areas are also hard to reach via cheap license-free solutions such as WiFi, as these technologies, operating in 2.4 or 5 GHz bands, have a very limited connectivity range.

In the 50-800 MHz band, a large block of UHF and VHF frequencies has recently been freed due to the analog to digital TV transition. This spectrum, called white spaces, promises to deliver an affordable means of providing wide area coverage. It is extremely attractive for rural areas as the propagation range is an order of magnitude higher than in the bands used by competing technologies. However, a distributed, resource-efficient solution for network organization, especially for spectrum allocation within a network, is needed for further proliferation of rural area white space deployments.

The goal of channel allocation is to assign one of the available channels to each of the network nodes. Traditionally, the issue has been expressed as the graph coloring problem where a color (channel) is assigned to a node so that network interference is minimized, and consequently the capacity is maximized. In a network operating over a small set of available frequencies, such as WiFi, channels generally do not exhibit significant differences in terms of propagation properties. White space networks, however, operate over a very large span of frequencies, and propagation properties can vary drastically over these channels. Channel assignment in such a network has to satisfy conflicting goals: maximize useful transmission by preferring channels with superior propagation properties, and minimize interference by favoring channels that propagate over a shorter radius.

Since we propose white space connectivity for impoverished regions, we concentrate on making our solution as cost efficient as possible. Therefore, we propose to reuse the existing TV antennas already installed in even the most remote rural areas. Unfortunately, this further complicates the problem of channel allocation as these antennas

exhibit uneven and unpredictable propagation behavior over the wide white space spectrum. Any analytical solution that provides a clear picture of frequency quality becomes impossible, and a direct inference of propagation properties is needed.

In this paper we successfully address the above challenges by designing a lightweight frequency profiling methodology to evaluate channel quality and a novel channel allocation method that assigns operating frequencies to base stations with the goal of minimizing the impact of interference over the useful signal levels in a network.

We compile these contributions into a practical channel profiling and allocation scheme for wide area white space networks called VillageLink. We test VillageLink's frequency probing mechanism on a long-distance software-defined radio white space link we deployed and confirm that antenna effects and the environment are a significant reason for high propagation diversity among white space channels. Through simulations we evaluate VillageLink's channel allocation. We show that our frequency-aware channel allocation leads to up to twice as much network capacity than an alternative heuristic based on interference avoidances. In addition, VillageLink preserves fairness among users even when the density of the network is high. With its high performance, efficient resource usage and distributed nature, VillageLink is a highly practical solution for wide area white space coverage in rural areas.

7.2 Wide-area White Space Networks

White spaces represent a historic opportunity to revolutionize wide area wireless networking. White spaces not only deliver much greater communication range than Gigahertz frequencies, they also support non-line of sight communication, including transmission through vegetation and small obstacles, which makes them highly suitable for various terrain configurations. However, white space networks have to deal with

unique peculiarities of transmission over a wide band of relatively low frequencies, and should enable license-free unplanned deployments in rural developing regions.

7.2.1 Wide band frequency selectivity

The variation in the free-space loss across a band is termed "dynamic range" and is calculated as follows:

$$D_{dB} = 20 \log (f_U / f_L)$$

where f_L and f_U are the lowest and the highest frequency in the band, respectively. In Table 7.1 we summarize the dynamic range of a number of traditional wireless systems. Free-space loss in a traditional wireless network, such as WiFi or GSM, is relatively uniform over the range of frequencies these networks operate on. The reason for low dynamic range in these networks lies in the fact that they either operate over a relatively narrow band of frequencies, such as 50 MHz for GSM and 80 MHz for 2.4 GHz WiFi, or they operate on high central frequencies where the difference between the lowest and the highest frequency diminishes, as is the case with 5 GHz WiFi. White spaces, however, operate on a wide band of low frequencies, and the difference in propagation between white space frequencies can be large. Note that the same issue does not arise in GSM (as well as 3G and 4G/LTE) networks, that can also operate on a wide range of frequencies (e.g. GSM850, GSM900, GSM1800). Unlike with white spaces, in these networks once the band selection is done the operation is restricted to a single relatively narrow range of channels.

Besides wide dynamic range, white space links often experience uneven fading due to antenna patterns and the environment. The fractional bandwidth (*FB*) for a frequency band, calculated as a ratio of operating bandwidth and the central frequency, determines how wideband an antenna should be in order to have the same gain over all frequencies with the band. From Table 7.1 we see that white spaces require significantly wider band

Table 7.1: Dynamic range and fractional bandwidth of different wireless systems.

Technology	f_L**(MHz)**	f_U**(MHz)**	***D*(dB)**	***FB*(%)**
802.11 (2.4 GHz)	2412	2484	0.26	2.9
802.11 (5 GHz)	5170	5700	0.85	9.8
GSM900	935	960	0.23	2.6
White spaces	43.25	797.25	25.31	179

antennas than GSM and WiFi. Such antennas are hard to design; high gain across the full frequency range being near impossible. Consequently, white space links are prone to the effects of imperfect antennas and surrounding structures.

To confirm this statement, we deployed a 3 km outdoor non-line-of-sight white space link. Each of the link nodes consists of a USRP2 radio and a dual core 2.4 GHz Pentium PC running GNUradio software. One node acts as a transmitter and sends probes at every 1 MHz over the white space spectrum. Another node, the receiver, scans the spectrum with 1 MHz spacing both with the transmitter turned off (baseline scan) and with the transmitter sending probes (signal scan).

Figure 7.1 shows the received signal strength across the UHF TV band. Three TV stations were detected and probes did not occur at these frequencies. The received signal strength does not fall of monotonically with increasing frequency, which would be the case if only free-space loss determined the propagation loss. Instead, due to the antenna characteristics, cabling and the environment, the propagation loss is non-uniform across the UHF band. Using the WIPL-D antenna modelling package we created a model of the deployed antenna. The results are shown in Figure 7.2. While an antenna with no surrounding structures has a more predictable gain pattern, when surrounding structures and antenna imperfections, such as bent or missing elements, are introduced, the antenna gain pattern has far less predictability. Residential TV Antenna installations often require long lengths of low-grade coaxial cable which can easily get dam-

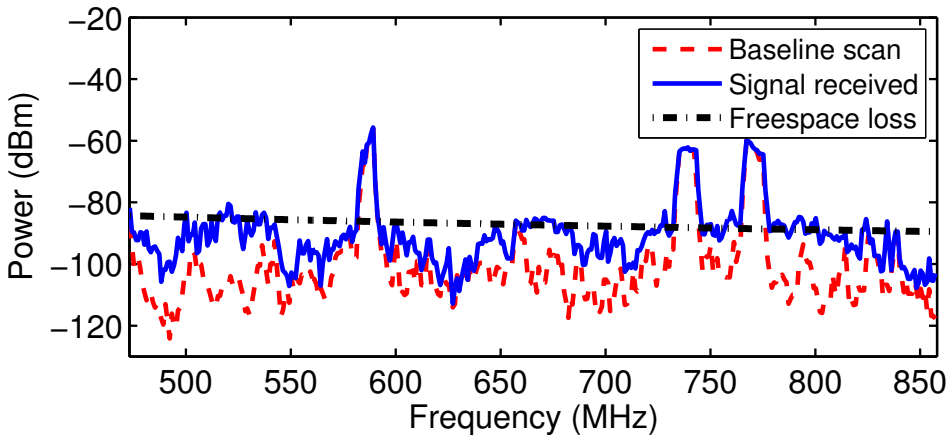


Figure 7.1: Analysis of received signal strength over the UHF band using 8 dBi Yagi antennas at transmitter and receiver. The plots demonstrate that received signal strength is difficult to predict as it is dependent on a complex mix of antenna gain patterns, cable issues and environmental structures around antennas and between the transmitter and receiver.

aged during installation by home users. Analysis of periodicity of dips in the received signal strength in Figure 7.1 indicate a cable that has been crushed or bent tightly approximately 2 m from the connector. Predicting the type of TV antenna being used, the structures surrounding the antenna or cable imperfections is not possible and provides strong support for frequency probing in white spaces.

In addition to antenna effects, a part of frequency selectivity may stem from the environment and terrain effects. *Shadowing*, i.e. slow fading due to physical obstacles on the signal path, would still be detected and accounted for with frequency probing. Unlike shadowing, *multipath*, which leads to rapid variation of propagation within a channel, cannot be captured through our current probing method. However, channel allocation only requires knowledge of the average channel gain of the channel that is captured by frequency probing.

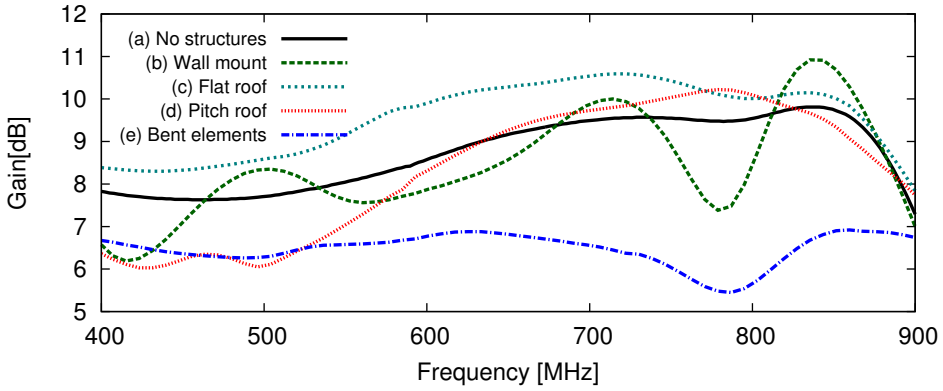


Figure 7.2: Simulated Antenna gain for double-yagi antenna used in outdoor white space link. The plot shows the following scenarios: (a) Antenna with no nearby structures; (b) antenna mounted on the side of a wall; (c) antenna mounted on a flat roof; (d) antenna mounted on a pitch roof; and (e) antenna with imperfections due to bent and lost elements.

7.2.2 Channel assignment in white space networks

The problem of channel assignment in wireless networks is often expressed with graph coloring, where each color represents a different channel. For a link, one of the available central frequencies is assigned so that a goal, such as maximum throughput, is achieved. In the channel allocation literature on traditional wireless networks all colors are considered equal in terms of their propagation properties [134]. However, in white spaces, the transmission range varies significantly among frequencies in the band due to the wide dynamic range and antenna effects (see Table 7.1 and Figure 7.2). Therefore, *selection of the operating frequency can impact the existence of a link itself.* This further complicates the problem of graph coloring, as now not all colors are equal. Figure 7.3 show one such example where a tradeoff between establishing links and avoiding interference is hard to achieve. In a white space network the color affects the graph structure, thus the existing approaches to frequency assignment are not directly applicable.

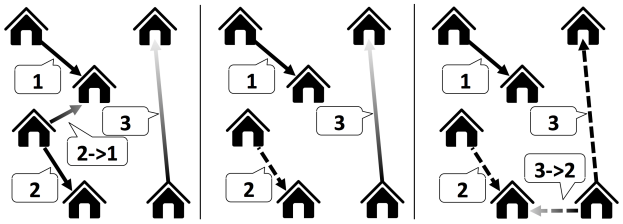


Figure 7.3: A simple example of the challenges of frequency assignment in white spaces. We want to establish links 1, 2 and 3; two channels represented by solid (inferior propagating frequency) and dashed (superior propagating frequency) coloring are given. In (a) all links operate on the inferior frequency; however there is interference from link 2 at link 1's receiver and link 3 is too long to be established on this frequency. In (b) the interference is resolved by switching the frequency for link 2. In (c) link 3 is established through assignment of the superior frequency. However, the superior frequency propagates further, thus interference at link 2's receiver is introduced.

7.2.3 Network Architecture

The network scenario that describes the setting in which VillageLink will operate is given in Figure 7.4. In this paper, we consider wide-area white space networks that consist of individual base stations (BSs), each with a set of associated customer-premise equipment (CPE) clients. We term one such BS with its CPEs *a cell*. A BS and all the CPEs within a cell operate on the same channel*; thus, when considering channel allocation we use "BS" and "cell" interchangeably. All cells that are operate within the same administration are called a WRAN domain. The existence of TV transmission and other, interfering, white space networks not in our service set reduces the number of channels available to the BSs within our domain. The aim of our work is to develop a channel allocation algorithm, as well as supporting structures such as a MAC layer

*We envision OFDMA channel sharing among the CPEs of a BS. Such an approach is mandated by the IEEE 802.16 and IEEE 802.22 standards. We leave the details of subcarrier allocation as the future work, and in this paper concentrate solely on channel allocation at the BS level.

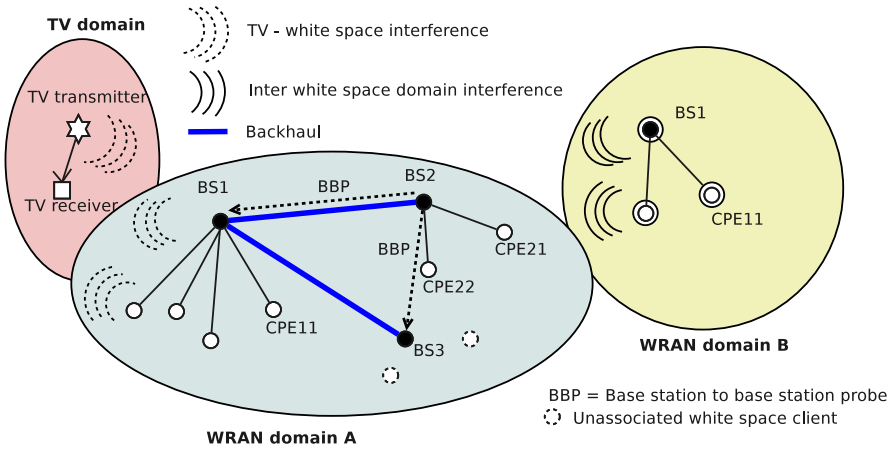


Figure 7.4: Layout of a targeted white space network showing interference scenarios between television and white spaces, and between white space networks in different domains. White space base stations within the same domain send base station to base station probes (BBPs) to calculate the channel conditions among themselves.

and a frequency probing mechanism, so that the overall network performance within our white space WRAN domain is maximized.

We assume that base stations are connected with a back channel. This can be another white space radio tuned to a common channel that does not interfere with the set of channels available for the base station-to-client connectivity. Moreover, because the amount of control data sent over the back channel is low, a VHF/UHF packet radio, or any other low bandwidth communication technology, can be used.

7.3 Channel Probing and Medium Access for Wide-Area Networks

Channel probing is an important tool for propagation evaluation over a wide white space frequency range. Unfortunately, the existing MAC protocols proposed for wide area networks [43, 118, 124] do not explicitly support frequency profiling. The MAC

protocol that most closely resembles our proposed system is IEEE 802.22. The protocol has built-in protection for primary users and mechanisms to move to new channels but has no built-in mechanism to choose from a set of available channels. It specifies that the channel may be chosen from the available list by an operator or by a "local routine". Instead of rebuilding an entire MAC layer, we propose to extend the 802.22 protocol to include a feature that performs frequency profiling on all available channels. In our frequency profiling scheme we measure the SNR of the probe that was received at a base station using the previously measured power spectral density of channel with no probes and the power spectral density measured when a probe is present. We assume that the network nodes are static.

In order to perform frequency profiling between base stations and between base stations and clients on all available channels, a mechanism is required to coordinate probing timing for channel probe senders and listeners. We use a token mechanism, in which a base station only sends channel probes when it has a token. All other base stations and clients without tokens are in a listening state. When a base station is in a probing state, it sends a probe at the beginning of each superframe. It sequentially steps through the full white space TV channel set and only sends a probe on non-interfering channels. A breath-first traversal of a spanning tree of the graph is used to ensure that a token traverses the graph of base stations connected by a back channel.

The 802.22 specification makes use of clients to sense for primary users and extend the sensing coverage area. We propose to use a similar notion when listening for probes. Clients of one base station experience interference from all other base stations. To account for this clients can be instructed to listen for probes from base stations with whom they are not associated. Frequency profiling results for clients are sent back to the associated base station on the final upstream frame once the client has listened on the full set of white space channels. The average SNR value of a received probe heard at a base

station and its associated clients is used to incorporate the average interference on the system. These SNR values, from each of the cells that received the probes, are unicast on the back channel to the sending base station. Results received at sending base stations are distributed to neighbouring base stations, where two base stations are defined as neighboring if a probe can be exchanged between them on at least one frequency.

Once the probing process is completed each base station stores a "localized set" of information on signal propagation at different frequencies: 1) within its own cell, obtained through aggregation of probing results from the cell's CPEs; 2) between itself and other base stations and associated CPEs that overheard probes; and 3) within cells that are served by each of the neighboring base stations. The number of channels in a white space network is in the order of few tens. Each BS performs a single round of probing, and it is only necessary to store a "localized set" of channel information (one float number per channel) due to the distributed nature of the channel allocation algorithm described in Section 7.4. Thus, we expect that the amount of information that has to travel via the back channel, in a sparse rural area network, remains in the order of a few kilobytes per BS.

7.3.1 Calculating probe SNR

Each probe is simply a pseudo-noise (PN) sequence modulated with DBPSK. The client calculates the average power received, $Pr_{avg}(c_i)$, measured over the probe listen window for each non-interfering channel, c_i.

$$Pr_{avg}(c_i) = \frac{1}{N}\sum_{n=1}^{N} s(n)^2 + CF$$

where $s(n)$ is the signal received, n is the sample number, N is the total number of samples and CF is the correction factor, which is calculated by calibrating the receiver.

Average power is a low complexity ($O(N)$) calculation and the cognitive radio can carry this out in real time using a cumulative average.

The SNR of the probe at each channel, c_i, with no TV interference can then be calculated by using a noise level N_{avg} from an initial scan when no probes are present.

$$SNR(c_i) = Pr_{avg}(c_i)/N_{avg}(c_i)$$

From the measured SNR we can extract the channel gain :

$$H(c_i) = \frac{SNR(c_i) \cdot N_0 \cdot W}{Pt} \tag{7.1}$$

where N_0, W and Pt denote the noise constant, channel width and the transmission power, respectively.

These SNR measurements are able to capture base station to client channel gains and channel gains (interference) between a base station and other cells that include the base station and associated clients. We do not capture client to client interference (client uplink effecting another client downlink on the same channel) as this would require clients to carry out probes and not allow our system to scale. Moreover, base station superframes are time-synchronized in 802.22 and only a small portion of the uplink frame is likely to overlap with the downlink frame of another cell. In addition, 802.22 is also able to move into a co-existence mode with an adjacent cell experiencing interference in which frames are fully coordinated between cells. We also assume channel reciprocity for our measurements – a common assumption for systems using the same channel for up and down links.

7.4 Channel Allocation

In this section we devise a distributed channel allocation algorithm that uses information obtained through frequency profiling (Section 7.3) and does not incur channel switching overhead typical for other allocation schemes. Our approach is based on the annealed Gibbs sampler, a technique that can help us minimize a target function in a distributed way. In the next subsection we present the basics of Gibbs sampling. An interested reader can find more details about Gibbs sampling in [30]. We then cast our problem to the Gibbsian framework and sketch the channel allocation algorithm.

7.4.1 Gibbs Sampling

The Gibbs sampler is a Markov chain Monte Carlo technique for obtaining random samples from a multivariate probability distribution. The sampler is useful in situations where the joint distribution is unknown or difficult to sample, but the conditional distributions of variables are known and easy to sample. In a nutshell, the Gibbs sampler that draws samples from a multivariate probability distribution $p(x_1, ..., x_N)$ works as follows:

- Initialize all variables $x_1, ..., x_N$ to (random) starting values $x_1^0, ..., x_N^0$.
- In every iteration $j = 1..k$, sample each variable x_i from the conditional distribution $p(x_i|x_1^j ..., x_{i-1}^j, x_{i+1}^{j-1}, ..., x_N^{j-1})$ to obtain x_i^j.

After the above process is finished, we are left with $x_1^j, ..., x_N^j; j \in [1..k]$ samples from the joint distribution p.

We can solve the channel allocation problem through Gibbs sampling, if we obtain the samples from a multivariate probability distribution that:

1. is related to overall network performance.

2. depends on the selected operating channel of each of the base stations.

3. isolates the impact of each of the base stations on the total optimization function.

4. can be calculated in a distributed way and sampled independently at each of the base stations.

5. favors states that lead to maximum network performance.

In the following section we develop a network performance metric that can be used as a basis for a probability distribution that satisfies the above demands.

7.4.2 Network Performance Metric

Traditionally, the goal of a channel allocation protocol is to assign available channels to BSs so that the total network capacity is maximized. The capacity $C_i(c_p)$ of a single cell operating on the channel c_i is:

$$C_i(c_i) = \sum_{k \in \mathbb{K}_i} W_k \log\left(1 + SINR_{ik}(c_i)\right)$$

where $\mathbb{K}_i$ is the set of CPEs within the cell, W_k is the width of a part of the channel c_i used by CPE k, and $SINR_{ik}(c_i)$ is the signal to interference plus noise ratio at the CPE k. We approximate the presence of all clients within the cell with a single *virtual CPE* with an SINR value $SINR_i(c_i) = \sum_k SINR_{ik}(c_i)/|\mathbb{K}|$. The cell capacity is now:

$$C_i(c_p) = W \log\left(1 + SINR_i(c_i)\right)$$

where W is the full channel width, essentially a sum of all W_k as a cell operates in an OFDMA mode. This approximation hides channel distribution within the cell and helps us concentrate on inter-cell interaction.

If we consider a network with N cells, with a given channel assignment $\mathbf{c} = (c_1, c_2, ..., c_N), c_i \in \mathbb{C}$, where $\mathbb{C}$ is the set of available channels, the total network throughput is a sum of all individual capacities at their respective allocated channels:

$$C(\mathbf{c}) = \sum_i C_i(c_i) = \sum_i W \log\left(1 + SINR_i(c_i)\right) \tag{7.2}$$

A single BS's decision on the operating channel changes the interference level at all its neighboring BSs. In the above equation the interference is accounted for in the SINR, which is embedded within the logarithmic function. Thus, the impact of a single BS on the total sum is hard to isolate, and the total capacity is not a suitable metric for distributed computation using Gibbs sampling. Centralized optimization using known polynomial complexity techniques, such as linear programming, is not directly applicable either, since the target sum involves non-linear factors and discrete variables.

One of the ways to circumvent this is to revert to a tighter problem formulation that prevents interfering base stations from concurrent transmission [92]. While this can be enforced in a network that employs carrier sensing and collision avoidance, in our setting long distances between base stations render such coordination inefficient [124]. In addition, allowing some interference often yields more capacity than restricting concurrent transmissions [104]. Another approach is to modify the optimization function and instead of maximizing capacity concentrate on minimizing total network interference [83, 123]. This approach is attractive for networks, such as WiFi, where these two goals are essentially interchangeable. In a white space setting, where available channels can differ drastically in terms of their propagation properties, *a channel allocation that leads to minimal interference may not necessarily lead to maximum capacity*.

We propose a novel network performance metric that allows distributed performance optimization with Gibbs sampling, and term it *cumulative interference plus noise to signal ratio (CINSR)*. It represents a sum of inverse of *SINR* experienced at

each of the cells. *CINSR* can be seen as the overall ratio of the impact of harmful factors, noise and interference, to the beneficial one, received signal strength. Thus, our goal is to minimize it:

$$CINSR(\mathbf{c}) = \sum_{i=1}^{N} \frac{1}{SINR_i(c_i)} \tag{7.3}$$

$$= \sum_{i=1}^{N} \frac{N_0 W + \sum\limits_{j=1..N}^{j \neq i} ch(i,j) PH_{ji}(c_i)}{PH_i(c_i)} \tag{7.4}$$

The first term in the numerator within the above sum is the thermal noise (a product of the channel width W and the noise constant N_0), whereas the second term is the sum of interference experienced at cell i, and originating from all other base stations that transmit at the same channel. Interference from a single source is a product of P - the transmission power and $H_{ji}(c_i)$ – the propagation gain from base station j to cell i on channel c_i. The function $ch(i,j)$ is equal to 1 if i and j operate on the same channel, and otherwise it is equal to 0. The denominator in the above equation is the average signal strength received by the clients of the BS i; the average channel gain from BS i towards the clients on channel c_i is denoted by $H_i(c_i)$.

We now isolate the impact of a single BS i on $CINSR(\mathbf{c})$ and term it *local CINSR*:

$$CINSR_i(\mathbf{c}) = \frac{N_0 W}{PH_i(c_i)} \tag{7.5}$$

$$+ \sum_{j \neq i} ch(i,j) \left(\frac{PH_{ij}(c_i)}{PH_i(c_i)} + \frac{PH_{ji}(c_i)}{PH_j(c_i)} \right) \tag{7.6}$$

Information needed for $CINSR_i(\mathbf{c})$ calculation, namely $PH_i(c_i)$, $PH_{ji}(c_i)$, $PH_{ij}(c_i)$ and $PH_j(c_i)$, is available locally at BS*i*, through channel probing described in Section 7.3.

7.4.3 The Gibbs distribution

The Gibbs distribution associated with the function *CINSR* and a positive temperature T is the probability distribution on c^N (the combined channel state space of all BSs) defined as:

$$\pi(\mathbf{c}) = \frac{e^{-\mathcal{CINSR}(\mathbf{c})/T}}{\sum_{\mathbf{c}' \in c^N} e^{-\mathcal{CINSR}(\mathbf{c}')/T}} \tag{7.7}$$

The above distribution is of a special interest as it favors states in which *CINSR* is low. In addition, the channel selected by BS*i* is independent of all non-neighboring BSs and the distribution fulfils all the conditions listed Section 7.4.1.

The Gibbs sampler draws a sequence of samples from the above distribution by having each of the BSs i independently sample its local Gibbs distribution $\pi_i(\mathbf{c})$:

$$\pi_i(\mathbf{c}) = \pi_i(c_i|\mathbf{c}_{\setminus i}) = \frac{e^{-\mathcal{CINSR}_i(c_i|\mathbf{c}_{\setminus i})/T}}{\sum_{\mathbf{c}' \in c^N} e^{-\mathcal{CINSR}_i(c'_i|\mathbf{c}_{\setminus i})/T}} \tag{7.8}$$

and transitions to the sampled local state, converging to the stationary distribution $\pi(\mathbf{c})$ (see Section 7.4.5). Here $\mathbf{c}_{\setminus i}$ denotes a fixed assignment of channels to all base stations but BS_i.

Distribution $\pi(\mathbf{c})$ favors low *CINSR* states when the temperature is low. While our goal is to minimize *CINSR*, by keeping the temperature low we risk getting stuck in a local minimum early in the process. *The annealed Gibbs sampler* introduces a slow decrease of temperature T to zero according to *a cooling schedule*. Therefore, in the beginning the probability of exploring a wide range of states is high, and as the time goes to infinity, the procedure converges to the minimum *CINSR* state (Proposition 3, Section 7.4.5). The choice of the schedule impacts the convergence speed, and we experiment with two commonly used schedules in the evaluation section.

7.4.4 Channel Allocation Algorithm

Algorithm 1 VillageLink channel allocation – distributed

1: {Executed at the base station i}
2: **while** $t < t_{end}$ **do**
3: $T = f(T_0, t)$ {f - schedule, T_0 - starting temperature}
4: **for all** channel $c_i' \in \mathbb{C}$ **do**
5: $\mathbf{c}' = (c_1, c_2, ..., c_i', ..., c_N)$
6: Calculate $CINSR_i(\mathbf{c}')$
7: **for all** channel $c_i' \in \mathbb{C}$ **do**
8: $\mathbf{c}' = (c_1, c_2, ..., c_i', ..., c_N)$
9: Calculate $\pi_i(\mathbf{c}')$
10: Sample a random variable according to the law π_i and choose the next channel of the BS i accordingly.
11: Send information about the newly selected channel to i's neighbors.
12: Switch the network interface to the last selected channel.

The Algorithm 1 is executed at each of the base stations. The temperature falls off with time, ensuring that the Gibbs sampler converges towards the global minimum of *CINSR*. The starting time for all the base stations has to be loosely aligned, and can be achieved through a standard synchronization scheme such as NTP.

Compared to some other distributed channel allocation schemes [83, 103], Algorithm 1 has an attractive property that no channel switching is needed until the convergence. To see why note that the calculation of the local *CINSR* is done after the probing process, and during the algorithm run the only variable parameter is $ch(i, j)$. At BS i this parameter can be updated irrespective of the actual operating channel of BS j. In every step a BS decides on its current channel and sends the decision to its neighbors, who then update their $ch(i, j)$ tables. Once the cooling schedule is completed base stations switch to their channel of choice (line 15 in Algorithm 1). This greatly speeds up the convergence, as the channel allocation process is not limited by the channel switching time.

7.4.5 Algorithm convergence

Convergence of a Gibbs sampler, and its annealed version, is a well researched topic. We refer readers to specialized texts about Monte-Carlo Markov Chains, such as [30] for details, and in this section show that the proposed method is a natural heuristic for solving the channel allocation problem.

Proposition 1. *The Gibbs distribution π (equation 7.7) represents a Markov random field.*

Proof. *A Gibbs potential* V associates a real number $V_\Gamma(\mathbf{s})$ with each subset Γ of a set S. The potential is determined by the state $\mathbf{s}$ of the nodes in Γ and is defined as zero if Γ is not a clique. *An energy function* $\mathcal{E}(\mathbf{s})$ maps each of the graph states to a real number. We say that the energy function *derives from the potential* V if:

$$\mathcal{E}(\mathbf{s}) = \sum_{\Gamma} V_\Gamma(\mathbf{s}) \tag{7.9}$$

where the summation goes over all subsets of the set S. The Gibbs distribution where the energy derives from a Gibbs potential is a Markov random field (pg. 260 in [30]), and we proceed with showing that the function that we use to construct the Gibbs distribution in equation 7.7 – *CINSR(c)* derives from the Gibbs potential.

We can represent *CINSR* as a sum of local impact of cliques of the graph of base stations $\mathcal{A}$. *CINSR* then takes the form described by equation 7.9 and can be used as

the energy function for Gibbs sampling:

$$\begin{aligned} CINSR(\mathbf{c}) &= \sum_{i \in \mathcal{A}} \frac{N_0 W}{PH_i(c_i)} + \\ &+ \sum_{\{i,j\} \in \mathcal{A}} ch(i,j) \left(\frac{PH_{ij}(c_i)}{PH_i(c_i)} + \frac{PH_{ji}(c_i)}{PH_j(c_i)} \right) \\ &= \sum_{\mathcal{B} \subset \mathcal{A}} V_{\mathcal{B}}(\mathbf{c}) \end{aligned}$$

Here V denotes the Gibbs potential. The potential is defined for all subsets $\mathcal{B}$ of the set of base stations $\mathcal{A}$ as:

$$V_{\mathcal{B}}(\mathbf{c}) = \begin{cases} N_0 W / PH_i(c_p) & \text{if } \mathcal{B} = \{i\} \\ ch(i,j) \left(\frac{PH_{ij}(c_i)}{PH_i(c_i)} + \frac{PH_{ji}(c_i)}{PH_j(c_i)} \right) & \text{if } \mathcal{B} = \{i,j\} \\ 0 & \text{if } |\mathcal{B}| \geq 3 \end{cases}$$

Note that the potential is non zero only for cliques of size one and two. Thus, energy $CINSR(\mathbf{c})$ derives from the Gibbs potential and, consequently π is a Markov random field. □

Proposition 2. *For a network of N base stations, each running a Gibbs sampler over its local Gibbs distribution $\pi_i(\mathbf{c})$, channel allocation converges in variation[†] towards the Gibbs distribution π.*

Proof. The process can be described as a Gibbs sampler on a finite state homogeneous Markov chain represented by the selected channel allocation, for which the Gibbs distribution (equation 7.7) is the invariant probability measure. Example 6.5, pg. 288 in [30] proves that such a sampler converges in variation to the target distribution, and that the convergence takes place with geometric speed. □

[†] *Convergence in variation describes convergence of an array of samples to a probability distribution and is defined in [30], pg. 128.*

Note that direct sampling of the capacity (equation 7.2) does not provide any guarantees on the performance as the capacity equation cannot be transformed to an energy function that derives from the Gibbs potential. Thus, we develop *CINSR*.

Proposition 3. *For a fixed network of N base stations implementing Algorithm 1, channel allocation converges in variation towards a limit distribution that only puts positive probability mass on the states of minimum global energy.*

Proof. The proof is analog to Example 8.8, pg. 311 in [30]. The annealed Gibbs sampler converges according to a strongly ergodic nonhomogeneous Markov chain: it converges in variation to a limit distribution that only puts positive probability mass on the states of minimum global energy. Conditions that the cooling schedule has to satisfy in order for convergence to happen can be found in [62]. □

7.5 Evaluation

The VillageLink system consists of our frequency profiling method built on top of the 802.22 MAC protocol, and the channel allocation algorithm based on Gibbs sampling. Experimental evaluation of such a system is challenging due to the need for a wide area outdoor deployment. In addition, off-the-shelf 802.22 equipment is not yet commercially available, and software defined radio platforms cannot support the synchronization that the MAC protocol requires [109]. Therefore, we evaluate our protocol in a simulated setting. However, the initial experimental investigation of channel probing and frequency selectivity in white spaces, presented in Section 7.2.1, was performed on a 3 km outdoor link.

7.5.1 Simulation Setup

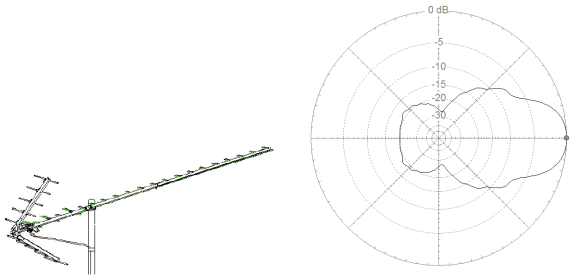
For a comprehensive evaluation of the channel allocation algorithm, we rely on a Matlab-based custom simulator. The simulator allows us to scale our experiments over a number of cells, and to model different network layouts. We explicitly take into account high variability of signal propagation in the white space band by modeling propagation with the Friss transmission equation:

$$P_r = P_t + G_t + G_r + 20 \log \left(\frac{\lambda}{4\pi R} \right)$$

where $\lambda, R, P_r, P_t, G_t$ and G_r are the wavelength, distance between antennas, received power, transmitted power, transmitter antenna gain, and receiver antenna gain, respectively. Antenna gains depend on specific devices used and their orientations. Earlier, in our outdoor testbed, we confirmed that frequency dependence of antenna gain is the most dominant factor that leads to the frequency diversity in white spaces (Figure 7.1, Section 7.2), thus we model antenna effects in detail.

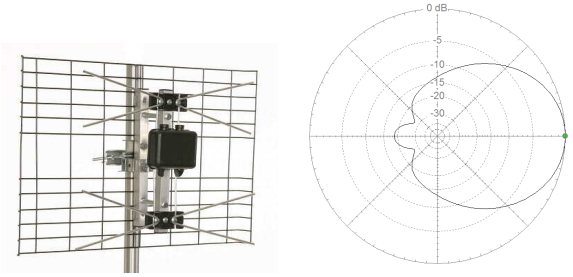
We use publicly available antenna models[‡] and the Numerical Electromagnetics Code[§] antenna modeling software to examine propagation over different frequencies with different antennas. Figures 7.5(b) and 7.6(b) show the radiation patterns seen from the the center frequency (598 MHz) of the white space band for two different antennas. In Figure 7.7 we plot frequency dependence of antenna gain. We found that the shape of the antenna pattern does not change significantly for different frequencies. The gain, on the other hand, changes significantly and unpredictably, as seen in Figure 7.7. Thus, in the simulations we use the antenna pattern shape of the center frequency to account for antenna orientation, and we use the full gain over frequency diversity.

[‡] www.hdtvprimer.com/ANTENNAS/comparing.html
[§] www.nec2.org



(a) (b)

Figure 7.5: Wineguard PR9032 UHF Yagi/corner reflector antenna used as a base station antenna in our evaluation. Showing (a) the antenna design and (b) its radiation pattern seen from the top of the antenna.



(a) (b)

Figure 7.6: AntennasDirect DB-2 2-Bay UHF antenna; one of the client antennas used for the evaluation. Showing the antenna in (a) and its radiation pattern seen from the top of the antenna in (b).

All base stations in our simulations use the Yagi antenna from Figure 7.5, as this antenna exhibits the best performance of all the antennas that were modeled. In our simulation we assume clients make use of existing TV antennas used to receive terrestrial TV broadcast signals. Operators have no control over the variety of antennas used by clients and we randomly select antennas from a set of 17 possible client TV antennas ranging from outdoor Yagi antennas with a gain of 15dBi to simple indoor loop antennas with a gain of 3dBi.

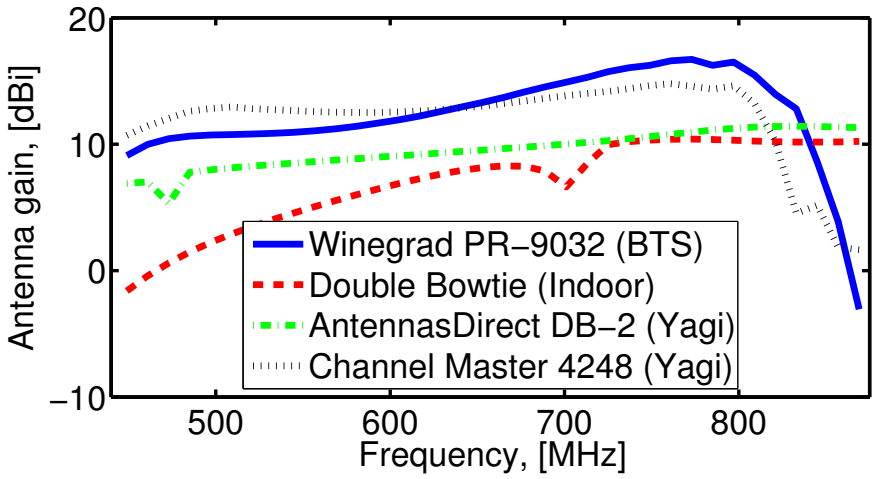


Figure 7.7: Antenna profiles of four of the antennas used in our evaluation. One of the profiles, Wineguard PR-9032, corresponds to the BS antenna; the other three correspond to client antennas.

We run our experiments over a white space band from 443 MHz to 875 MHz as the antenna models we use perform reasonably well within this range. The band is divided into 36 TV channels, each 6 MHz wide, with a 6 MHz guard band between adjacent channels. In all the experiments we simulate a $100km \times 100km$ field with random BS placement and random antenna orientation. Each of the BSs has a single associated virtual client at a distance uniformly picked from 0.2 km to 20 km and with its antenna pointed directly towards the BS. We also simulate a TV station that covers a part of the field with its signal and occupies two adjacent channels.

7.5.2 Channel Allocation Convergence

Our solution to channel allocation represents a heuristic as we cannot give guarantees on how long will it take for the MCMC process to reach the target invariant distribution (i.e. we do not know the *mixing time*). To gauge the practical behavior we simulate *Algorithm 1* in a network of ten base stations and five white space channels

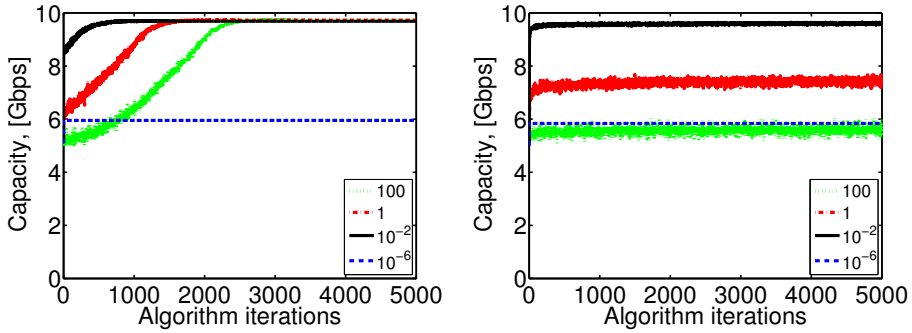


Figure 7.8: Algorithm convergence with the (a) exponential, and (b) logarithmic cooling schedule. Each line corresponds to a different starting temperature.

that are available for communication. We are interested in the algorithm convergence under different Gibbs sampling parameters. We experiment with two common cooling schedules: *a)* logarithmic: $T = T_0/\log(t+2)$, proposed in [57] and *b)* $T = T_0\alpha^t$, proposed in [86]. Here T denotes the temperature at time t, T_0 is the starting temperature, and $\alpha \in [0, 1]$ is a real number; we empirically find $\alpha = 0.995$ to work well in our experiments.

The selection of the starting temperature is important for proper annealing. In Figure 7.8 we plot total network capacity achieved with the two schedules and four different starting temperatures for each. Each point in the graphs is an average over 100 runs. The impact of the starting temperature is clearly visible: the higher T_0 is, the more time it takes for the algorithm to converge. At the same time, higher temperatures ensure exploration of a large part of the solution space, and generally lead to a better solution. We can also see that $T_0 = 10^{-6}$ does not result in any variation of capacity as the algorithm progresses – the sampler is "frozen" and BSs will stick to the initial channel allocation without exploring the full solution space. There is a trade-off, dictated by the starting temperature, between the convergence time and the assurance that the optimal

value will be found. In the rest of the paper we fix T_0 to 1, a value that allows full exploration of the solution space and converges in a reasonable amount of time.

We observe much faster convergence with the exponential schedule, which converged in all but one case ($T_0 = 10^{-6}$). The logarithmic schedule did not converge in 5000 iterations for $T_0 = 100$ and $T_0 = 1$. In the rest of this section we rely exclusively on the exponential schedule.

7.5.3 CINSR as a Performance Metric

To confirm that *CINSR* is a good choice for the network performance metric, we compare it with an alternative – overall interference and noise in the network – which is often used as a metric in channel allocation algorithms [83, 123].

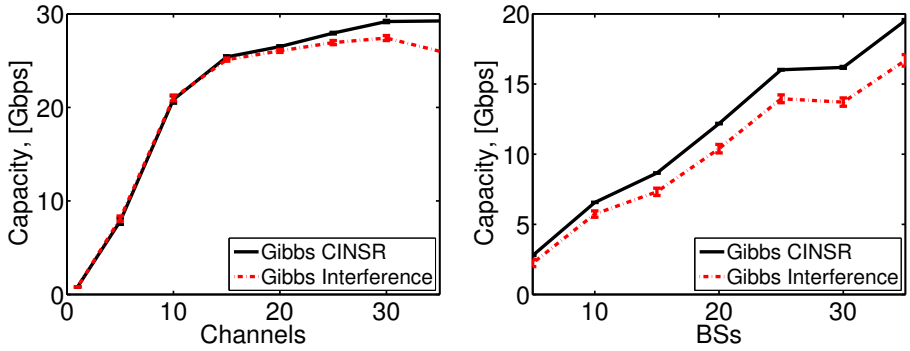
The total network interference and noise is defined as:

$$I(\mathbf{c}) = \sum_{i=1}^{N} \left(N_0 W + \sum_{j=1..N}^{j \neq i} ch(i,j) PH_{ij}(c_p) \right) \tag{7.10}$$

The impact of a single BS on the sum is defined as *the local interference*:

$$I_i(\mathbf{c}) = N_0 W + \sum_{j=1..N}^{j \neq i} ch(i,j) \left(PH_{ij}(c_p) + PH_{ji}(c_p) \right) \tag{7.11}$$

We modify equation 7.7 and equation 7.8 to include $I_i(\mathbf{c})$ instead of $CINSR_i(\mathbf{c})$, and $I_i(\mathbf{c})$ instead of $CINSR_i(\mathbf{c})$, respectively. The necessary conditions for the Gibbs sampler convergence still hold, and we apply an algorithm analogous to Algorithm 1. Note that, defined this way, the interference function still uses the results of channel probing, yet it does not account for the balance between well propagating channels that are preferred by the CPEs and inferior channels that minimize inter-cell interference.



(a) Channel under-provisioning. (b) Channel over-provisioning.

Figure 7.9: Comparison of the total network capacity achieved with CINSR and Interference metrics. We simulate under-provisioned and over-provisioned number of channels with respect to the number of base stations in the network.

7.5.3.1 Channel under-provisioning

We simulate a network with a number of contending BSs higher than the number of available channels, a typical case in the urban developed world. We put 50 cells in the same $100km \times 100km$ region. We experiment with a varying number of available channels. The total network capacity is plotted in Figure 7.9(a). Each point represents an average value of 20 runs of the algorithm with a different metric, *Gibbs CINSR* or *Gibbs Interference*, over the same topology.

When multiple cells operate on the same frequency the network is in a low SINR mode, and capacity can be increased by interference minimization. From Figure 7.9(a) we see that the two versions of the Gibbs sampler perform equally well with a small number of available channels. As we increase the amount of available spectrum, BSs have more freedom to operate at different channels with minimal interference. Therefore, frequency-dependent performance of CPEs associated with the BSs becomes an important factor that impacts total capacity. Since this factor is not accounted for in

equation 7.10, this version of the Gibbs sampler results in a channel allocation that delivers less capacity than the version that uses *CINSR*.

7.5.3.2 Channel over-provisioning

We now fix the number of available channels to 36 and compare the performance of the two versions of the algorithm with the number of BSs varying from 5 to 35. The total network capacity is plotted in Figure 7.9(b). Each point represents an average value of 20 runs of the algorithm (*Gibbs CINSR* or *Gibbs Interference*) over the same topology.

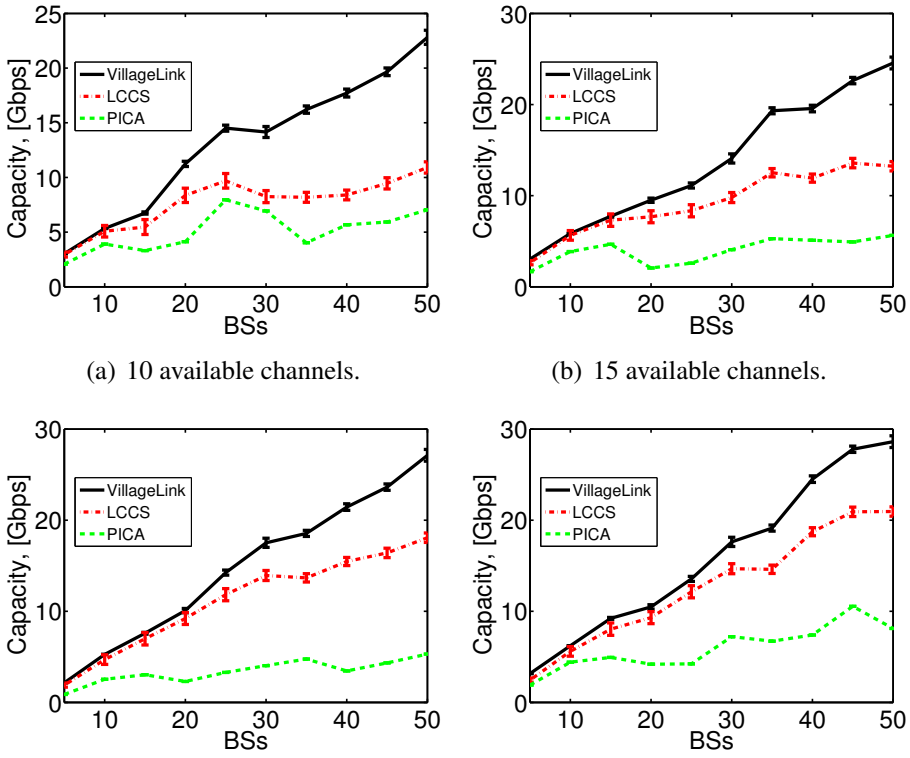
When the number of channels is greater than the number of BSs there is more than one allocation that leads to minimal interference. However, not all of the allocations are favored by the CPEs. Through the factor $H_i(c_p)$ CINSR accounts for the frequency dependent intra-cell preferences, and assigns channels that maximize capacity within each of the cells. The results presented here point out that *channel allocation in white spaces remains important even in rural areas where the channel availability is high* [63].

7.5.4 Comparison to alternative channel allocation methods

Channel allocation is a difficult problem to solve in a distributed setting. Heuristics are often used instead of a rigorous solution and we compare our approach with:

- *Least congested channel search (LCCS)* - a heuristic where each of the BSs individually scans for a channel with the least number of other BSs assigned to it [103].

- *Preferred intra-cell channel allocation (PICA)* - in this greedy method each of the BSs selects the channel for which it observes the highest channel gain towards its own CPEs ($\arg\max_{c_p} H_i(c_p)$).



(a) 10 available channels.

(b) 15 available channels.

(c) 20 available channels.

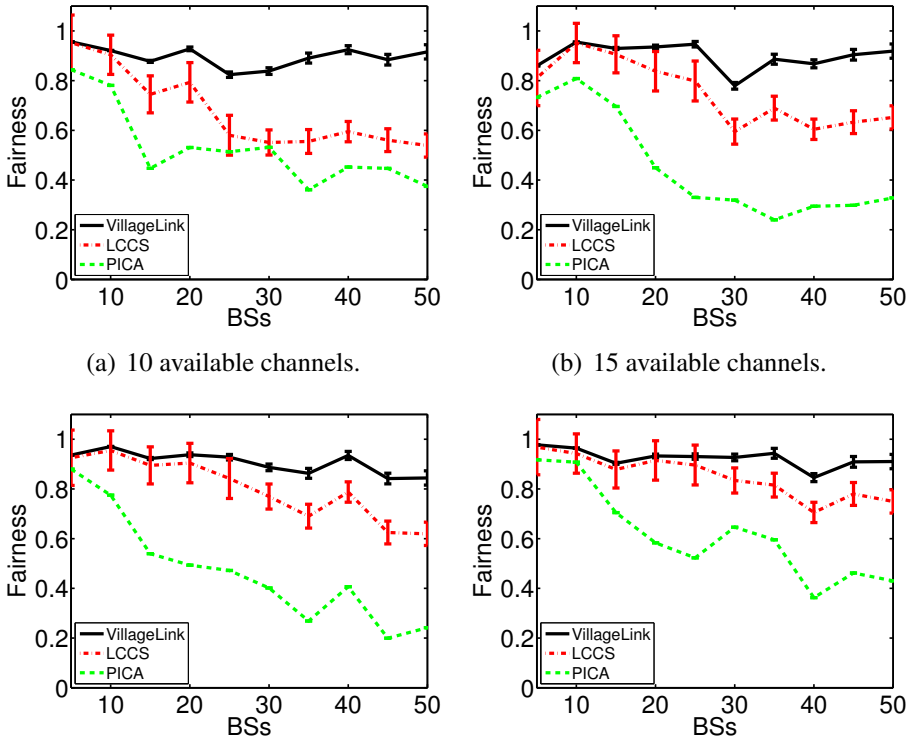
(d) 25 available channels.

Figure 7.10: Total network capacity with varying number of channels and base stations.

These heuristics optimize a non-submodular capacity function in a greedy manner, therefore may settle for a solution that is arbitrarily far from the optimal. VillageLink's convergence to the states of minimum *CINSR* is proven in Section 7.4.5. We compare the experimental behavior of different solutions in a number of scenarios encompassing various numbers of BSs and available white space channels. We run each of the algorithms 100 times in each of the scenarios.

7.5.4.1 Total network capacity

In Figure 7.10 we plot the total network capacity as we increase the number of cells in the system from 5 to 50. To ensure consistency among points in the graph, we do not



(a) 10 available channels.

(b) 15 available channels.

(c) 20 available channels.

(d) 25 available channels.

Figure 7.11: Fairness with varying number of channels and base stations (the closer the fairness index value is to one - the better).

generate a new topology every time we increase the number of cells, but add randomly placed cells to the existing topology. Each of the topology sequences are evaluated in environments with 10, 15, 20 and 25 available channels. We plot average values and two standard deviations (represented by error bars) for each data point.

VillageLink performs better or equal to the alternatives in all scenarios. The benefits of frequency-probing based channel allocation grow with the number of cells. In some scenarios, such as 50 BSs - 10 channels and 50 BSs - 15 channels, VillageLink delivers twice as much capacity as the next best alternative, LCCS. A comprehensive comparison of LCCS and PICA could unravel the importance of two conflicting goals

in channel allocation: minimizing interference and maximizing intra-cell capacity in isolation, and is left for future work.

7.5.4.2 Fairness

In Figure 7.11 we plot the Jain fairness index [74] for cell capacity with channel allocations determined by VillageLink, LCCS and PICA. We plot average values and two standard deviations (represented by error bars) for each data point. Although we designed VillageLink as a method to optimize total network capacity, it also ensures a remarkably fair allocation of resources. As the number of cells grows, the fairness of VillageLink is more pronounced as it stays close to 1 while the fairness indices of PICA and LCCS drop.

7.6 Related Work

Over the past decade most efforts to provide broadband connectivity to remote rural are based on modified WiFi [118, 124]. Propagation in White spaces bands are drastically different to WiFi bands and networking protocols need be reconsidered for new white space spectrum. In VillageLink we embrace 802.22 [43], a standardized protocol for wide area white space coverage, and augment it with novel channel probing and operating frequency selection mechanisms.

Channel assignment in wireless networks is traditionally treated as an NP-hard graph coloring problem [103] and numerous heuristic solutions have been proposed ([134] and references therein). Ma and Tsang [92] use an integer linear programming solution for frequency allocation to deal with channel heterogeneity in wide bands. However, frequency reuse is restricted to well defined interference domains where no two BSs are allowed to transmit at the same time. Motivated by [104], we rely on a

more sophisticated representation of interference, measuring its impact through probing to account for interference during the allocation process. Li and Gross in [91] propose a game-theoretic approach for distributed channel allocation in white spaces networks that considers transmission power and interference. This work, however, does not address the unique challenges brought by channel allocation in a large dynamic range.

Gibbs sampling was proposed in 1984 for image manipulation [57] Recently, its applicability to distributed channel allocation, client association and power control has been examined [83, 100]. VillageLink differs from previous approaches by incorporating frequency dependence of both useful signal transmissions and interference. In addition, VillageLink minimizes channel switching and information exchange among nodes.

7.7 Conclusion

White space networks are largely unexplored, and their straightforward implementation might prove difficult due to unique characteristics they exhibit. In this work we show how the heterogeneity of white space frequencies imposes unique challenges when it comes to channel allocation in a wireless network. Rather than simply minimizing interference, a channel allocation policy has to account for transmission quality over different channels as well. To tackle the problem we develop VillageLink, a channel allocation protocol that relies on the knowledge of signal propagation in the whole white space band before it performs distributed channel assignment that converges towards a network-wide optimum. Our work examines only one aspect of network adaptation. The complex nature of signal propagation over a wide frequency band opens up new possibilities for protocol design and further refinement of channel access in white spaces.

7.8 Acknowledgements

The material in this chapter is based on joint work with Veljko Pejovic, David L. Johnson, Elizabeth M. Belding and Albert Lysko.

Chapter 8

Connectivity: TxMiner – Identifying Transmitters in Real World Spectrum Measurements

8.1 Introduction

Spectrum is often monitored by spectrum regulators and licensed users. Spectrum regulators respond to complaints and identify unauthorized users of the spectrum. Licensees monitor the spectrum to identify utilization, and future planning. Existing work provides tools to identify known transmitters, through prior knowledge of time or frequency signatures.

However, a key missing capability in prior work is the ability to identify and characterize transmitters that do not have known signatures. For example, in Figure 8.1, it is very difficult to answer the following questions: Is the spectrum mask gathered over a time period from one, or more transmitters? What is the bandwidth of the multiple transmitters? What are the modulations schemes, and whether the transmitters are stationary or mobile? Are they frequency hopping? We attempt to answer many of these questions using TxMiner.

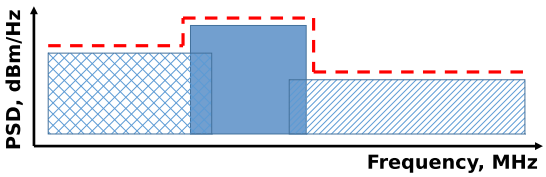


Figure 8.1: Example of overlapping transmitters.

The capability to characterize transmitters is useful in many scenarios. Regulatory agencies can learn how different parts of the spectrum are being used, and whether any part of the spectrum should be repurposed, or allowed for dynamic spectrum access [4, 7]. Second, spectrum licensees want to know if someone is illegally using their spectrum. These intentional, or in some cases unintentional, squatters (such as neon lights [13]), can cause unexpected interference to the licensee's operations. Third, for several military applications, the respective authorities might want to know about transmitters that use custom coding and modulations schemes, and whether the transmitters are stationary or mobile.

However, identifying these transmitters is non-trivial, as we show in Figure 8.1. The max-hold of spectrum might be dominated by the most powerful transmitter, thereby making it difficult to identify other transmitters. Furthermore, the fluctuations in the received signal makes it difficult to differentiate signal variations from one transmitter to signals from different transmitters. Moreover, even when temporal signatures are known for certain transmitters, it might be difficult to get sufficient time samples from wideband sweep based spectrum analysers.

In this paper we propose a system, called TxMiner, that identifies transmitters and their characteristics in raw spectrum measurements. It does so by applying machine learning techniques on measured spectrum data. TxMiner is based on a simple observation that wireless fading is log-normal [58], and that log of received signal strengths from the same transmitter will fall along the same gaussian curve. Therefore, teasing

apart different transmitters from spectrum measurements translates to identifying the corresponding gaussian distributions in the log of power values. We accomplish this using a novel use of Gaussian Mixture Models with Belief Propagation. We note that even though the curves might not always be Gaussian (due to shadowing, or multi-path), our machine learning approach is able to identify them with high probability. TxMiner is also able to identify various patterns, such as mobile transmitters, frequency hopping patterns, and multiple transmitters, as shown in the next section.

TxMiner takes as input the time-frequency power measurements over a period of time, applies the machine learning algorithms, and outputs a list of transmitters, and their characteristics – bandwidth, center frequency, and mobility. We have evaluated TxMiner on spectrum measurements from 30 MHz to 6 GHz from 4 locations, along with several controlled experiments, and found that it can accurately identify transmitters of different types: Wi-Fi, Bluetooth, WiMax, TV & FM broadcasts, as well as proprietary protocols. In most cases, it can accurately detect the bandwidth of these transmissions, and can also identify them even when these devices operate in overlapping bandwidths. To the best of our knowledge, TxMiner is the first system that can untangle transmitters from a mixture of wideband spectrum measurements.

8.2 Problem Definition

There are two essential components to the problem: (i) spectrum occupants (ii) spectrum sensors.

Spectrum occupants are wireless devices that operate in a given frequency range at a particular location. Those can be (i) primary incumbents, (ii) secondary devices (e.g. Dynamic Spectrum Access devices) and (iii) non-standardized transmitters such as military, foreign (e.g. close to international borders), and malicious transmitters.

Spectrum sensors are devices that continuously sense the spectrum and collect data about spectrum occupancy. Spectrum analyzers are typically sweep-based, and scan the entire spectrum of interest with a low resolution bandwidth (RBW). These sensors contribute geo-tagged spectrum traces in time-frequency domain with time resolution of Δt and frequency resolution of Δf. Each measurement point for a given time and frequency is a power level, P. Sometimes, dedicated spectrum analyzers could be installed that do not sweep, but instead continuously monitor a given portion of the spectrum.

Goals

The goal of TxMiner is to identify transmitters given spectrum power data over time and frequency. Specifically, we are interested in characteristics such as:

- center frequency of the transmitter
- bandwidth of the transmitter: even when they overlap
- number of transmitters
- whether the transmitters are mobile

We note that the goals are ambitious, and although TxMiner makes a significant advance at solving this problem, it still has some limitations. We will discuss those in detail in Section 8.7.

Applications of TxMiner

Since TxMiner can identify transmitters without known signatures, it enables several new applications. We list some of them in this section.

Mapping Spectrum Occupancy: The recent proliferation of mobile devices has significantly increased the demand for wireless spectrum. Consequently, spectrum regulators worldwide are evaluating usage of their spectrum in an attempt to repurpose unused spectrum for mobile connectivity. For example, the FCC has been mandated by Congress to create a spectrum usage map for all of US as part of the Spectrum Inventory Bill [4]. The problem is slightly different in developing countries where spectrum regulators frequently do not know about transmitters using different parts of the spectrum.*

Identifying Rogue Transmitters: Spectrum licensees, or in some cases the government, want to know of spectrum squatters. These are transmitters that, knowingly or unknowingly, illegally operate in the licensee's spectrum, thereby causing interference to the licensee's operation. Since the rogue transmitter's operations are mixed with the concurrent operations of the licensed transmitters, it is very difficult to identify them in a typical spectrum mask. For example, it took a team of several specialists more than 6 months to isolate the cause of interference to a cellular base station to be a neon light [13]. TxMiner can identify transmitters that do not match the pattern of the licensed operations. Many such instances of TxMiner can then be used to localize the rogue transmitters.

Characterizing Spectrum Squatters: A class of users, such as the military or spectrum regulators, would also want to know properties of the rogue transmitters, such as the bandwidth of operation, whether they are mobile or stationary, etc. TxMiner is able to mine this information from spectrum traces.

White Spaces Outside of TV Bands: Unused spectrum can be opportunistically used for data communication as long as it does not interfere with the primary users of the spectrum. The TV white spaces is one such example of this technology. However, ap-

*The Kenyan and Philippines government have requested help in identifying transmitters using parts of their spectrum [7].

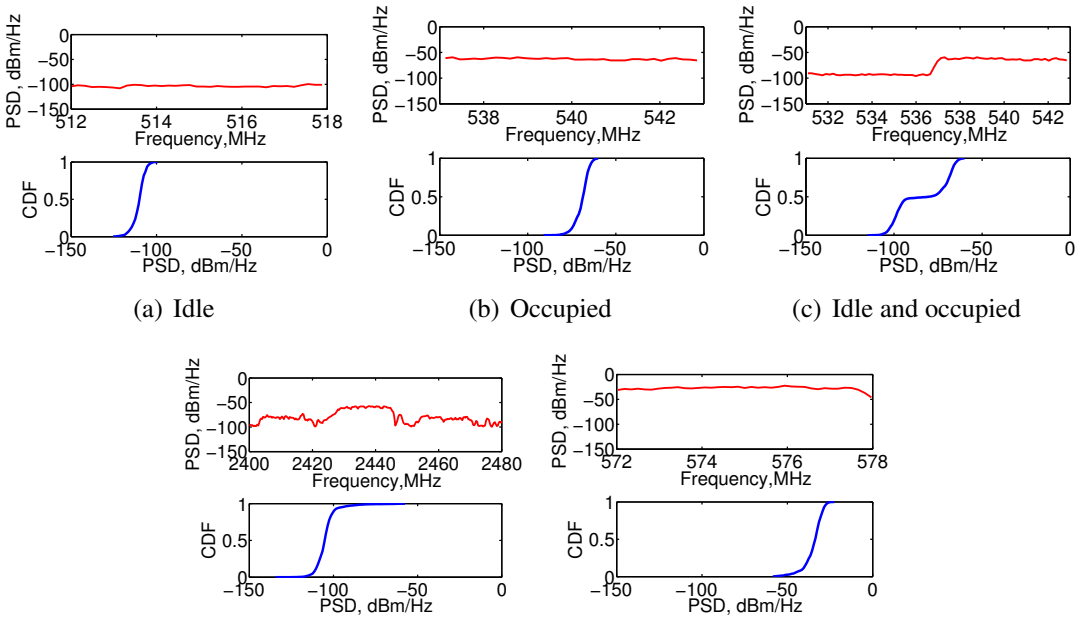
plying the same concept to other parts of the spectrum is non-trivial. The locations of primary users and their transmission characteristics might not be known to use a modelling based DB approach currently used for the TV white spaces. These problems can be overcome by spectrum analyzers running TxMiner, which can dynamically identify primary users and upload this information to the database.

Spectrum Management: Unlicensed spectrum can get extremely crowded, and is often used by new technologies that use custom physical and MAC layer protocols. For example, four device vendors that operate in the TV white spaces (Adaptrum, 6Harmonics, Carlson, Redline) each use their own PHY and MAC scheme. To add to this variety there are TVs and wireless MICs, and other transmitters. A network operator planning to use the white spaces would want to answer questions like: (i) how many types of transmitters are using the channel? (ii) how many transmitters of each type are present? (iii) what is the noise floor of the channel when these transmissions are not present?

A similar problem exists for the unlicensed 2.4 and 5 GHz of spectrum. New transmitters are licensed frequently, such as Amimon's wireless display, baby monitors, digital wireless MICs, etc. It is difficult to keep track of all the new types of frequency signatures. Even if that were feasible, it is very difficult to learn of the number of active transmitters only through spectrum measurements.

8.3 Methodology

In this section we first present the insights and theory behind the design of TxMiner. We then explain the algorithms and techniques in detail in the rest of the section.



(a) Idle (b) Occupied (c) Idle and occupied

(d) Frequency hopping (Bluetooth) (e) Mobile

Figure 8.2: Probability Distributions of Power Spectral Density for different occupancy scenarios.

8.3.1 Key Insights

The key insight behind TxMiner is that the *probability distributions* of measured Power Spectral Density (PSD) reveal a lot about channel occupancy. Consider a simple scenario when two static transmitters transmit in the same frequency band, but only a single transmitter is active at any given time. At the sensor, the received PSD level exhibits a log-normal distribution if no line-of-sight components are present and a Ricean distribution if there is a line-of-sight component in addition to the non-line-of sight components [58]. The resulting probability distribution as a result of both the transmitters being present is a mixture of the probability distributions of each transmitter, thereby resulting in a bimodal distribution. Furthermore, the logarithm of PSD level of each transmitter is Gaussian for log-normal distribution and has the shape similar to a shifted Gaussian for Ricean distribution. Therefore, the bimodal distribution is a mixture of two different Gaussian distributions when the two transmitters can be separated.

To illustrate this observation we study the probability distributions of different spectrum occupancy scenarios. We analyse the differences in distributions between wide and short-range transmissions as well as between mobile and stationary transmissions. Particularly, we focus on three usage settings: (i) static observation of a wide-range TV broadcast service, (ii) static observation of a short-range frequency-hopping transmission (e.g. Bluetooth) and (iii) transmission from a wide-range mobile transmitter.

Figure 8.2 presents the probability distributions for the studied spectrum occupancy scenarios. The top graphs present a max-hold of PSD over a time window of 100 seconds, while the bottom graphs present the CDF of all values measured in this window. We see that the distributions of one occupied and one idle TV channel (Figure 8.2(a) and 8.2(b)) are very similar in shape, however, the mean of the occupied channel is higher than that of the idle channel. In a frequency band, which is in part occupied and in part idle (Figure 8.2(c)), the probability distribution we observe is bimodal, reflecting

on the two spectrum activities. The means of the two modes correspond to the mean received power levels during the spectrum measurements. Next, Figure 8.2(d) presents the distribution of a Bluetooth transmission. Due to the short range of such transmissions, signal levels fluctuate less (e.g. effects such as multipath are not as pronounced as in wide-range transmissions), thus we observe a steep distribution centered around the noise floor of about -104 dBm with a tail at higher power levels that corresponds to the occasional frequency-hopping transmissions. Finally, in Figure 8.2(e) we present a distribution of a mobile transmitter. The dispersion of this distribution is larger than that of the static transmitter scenarios, which corresponds to the signal deviation when the distance between the mobile transmitter and the spectrum sensor is changing.

We harness these observations and design a methodology that makes use of machine learning approaches to determine the characteristics of the transmitters occupying particular spectrum range. We now give more detail about the machine learning techniques utilized in this work.

8.3.2 Harnessing probability distributions

In order to extract probability distributions from raw spectrum measurements, we resort to unsupervised machine learning. Intuitively, a Gaussian mixture model is a probabilistic model for representing probability distribution of observations in the overall population using an additive combination of Gaussians. We choose this approach for two reasons. First, since wireless signals from each transmitter undergo log-normal fading, the log-scale (dBm) measurements of power will follow a normal (Gaussian) distribution for each transmitter. Second, a key feature of Gaussian mixture models is that it can be used to model the presence of subpopulations within an overall population, without explicitly identifying the sub-population to which an individual data point belongs.

Gaussian Mixture Modeling has been a fairly successful and widely used unsupervised machine learning technique [51]. The GMM is a representation of any probability distribution in terms of weighted sum of individual Gaussians. Each Gaussian component in the mixture model is characterized via the mean and the variance and represents a sub-population. Furthermore, each of the components is associated with a weight and characterizes its contribution to the probability density. These weights, means and variances for each of the components comprise the parameters of GMM, which in turn are discovered via the Expectation-Maximization (EM) procedure that aims to find parameters that maximize the likelihood of the observed data. In addition to the parameters, one of the key by-products of the EM procedure is probabilities associated with each data point that reflect how likely that point arose due each Gaussian component. This is particularly important to deal with variations in the wireless signal due to multipath, mobility and shadowing.

In our problem setting, we consider all the entries of the PSD matrix and our proposal is to fit a Gaussian Mixture Models (GMM) to the entire observation set. The different components of the mixture model correspond to different transmitters and the useful by-product of the EM algorithm will automatically associate each observation of the PSD matrix with the different sub-populations (transmitters in our case).

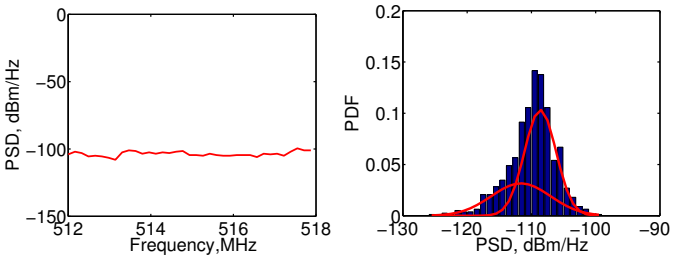


Figure 8.3: Example of fitting a model. [Left] Max-hold of PSD over time in channel 21 (512MHz – 518MHz) and [Right] fitted Gaussian Mixture Model.

Ideally, each fitted component should correspond to one transmission state, however, in reality this is not the case for several reasons. First, due to the signal deviation in wireless transmissions a single transmission event is often times represented by more than one Gaussian components. For example in Figure 8.3 we observe a single TV broadcast transmitter, however due to signal variation over time, a GMM method extracts two sub-populations from the channel measurements: one population corresponding to the actual transmission and another population corresponding to the noise. Second, in case of multiple simultaneous transmissions represented by the same component, we would not be able to differentiate between these transmissions. Lastly, while GMM can give us intuition about what might be happening in the spectrum band of interest, it does not allow for direct extraction of frequency and temporal transmitter characteristics. For these reasons, we cannot employ the model components directly into determining transmitter characteristics. Rather, we use the model components to get an intuition about what events might be occurring in the collected PSD data and use this information in subsequent analysis of transmitter characteristics. We differentiate between two types of GMM: non-informed and informed. In the non-informed case the only input to the algorithm is the PSD data and a guess about the number of components. In the informed case, we can additionally supply information about expected means and variances of components. Such guesses can be, for example, the expected noise floor. Supplying knowledge to GMM helps mitigate the effects of overfitting and leads to better representation of the data. In TxMiner we use both non-informed and informed GMM.

While GMM is a very effective tool for inferring the occupancy state, it does not exploit the spatial coherence in the frequency and the time domain. Specifically, in the PSD matrix it is easy to see that the adjacent cells (both in time and frequency) are more likely to see similar observations. Thus, unless the observations differ quite

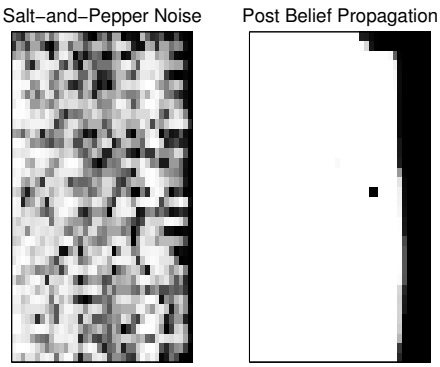


Figure 8.4: Example of loopy Belief Propagation.

significantly, there is no reason to consider different occupancy states between two neighboring cells. Since the GMM does not ensure such spatial regularization, it suffers from 'salt and pepper' noise, where adjacent cells often get assigned to separate occupance states (see Figure 8.4 (Left)).

We propose to alleviate this problem via spatial regularization using machine learning techniques popular in image segmentation literature [22]. In particular, we formulate an energy minimization problem where we consider adjacent cells in the PSD matrix (both in frequency and time) as neighbors. The goal of the energy minimization problem is to determine a solution that not only is aligned with the GMM results available from the previous step, but also is spatially smooth. Formally, we consider the following form of the energy:

$$E = \sum_{i} -\log p_{GMM}(s_i) + \sum_{ij} V(s_i, s_j, o_i, o_j). \tag{8.1}$$

Here, $p_{GMM}(s_i)$ is a unary term and simply depends upon the output posterior probabilities from the GMM. Intuitively, this term favors assignments that are obtained from the inference when fitting the GMM. The second term considers all pairs of neighbors

(i and j), and enables spatial smoothness by using a function $V(\cdot)$ that depends upon the corresponding observations o_i and o_j in the PSD matrix:

$$V(s_i, s_j, o_i, o_j) = \begin{cases} -loge^{-\beta|o_i - o_j|} & \text{if } s_i = s_j \\ -log[1 - e^{-\beta|o_i - o_j|}] & \text{Otherwise} \end{cases}$$

Note that the pairwise terms favors similar assignments to s_i and s_j only when the values o_i and o_j are similar. Intuitively, the pairwise term will favor dissimilar assignments to adjacent cells only when there is a large difference in observations in the PSD matrix.

An assignment that minimizes the above energy would aim to provide a solution that is spatially coherent as well as mostly aligned with the solution provided from the GMM procedure. However, determining the minimum energy assignment for such energies has been determined to be NP-complete in general. Reasonable approximation can be computed via message passing schemes such as loopy Belief Propagation [151]. In this paper we specifically, use the sum-product version of loopy belief propagation, where given the GMM inferences we formulate the energy and obtain a solution via loopy message passing until convergence. Figure 8.4 (Right) shows the result after running the loopy BP and we can observe that the resulting solution is spatially more regularized and does not have salt pepper noises.

8.4 Mining transmitters

This section gives detail into our methodology for extracting transmitters. We start by fitting the parameters for the Gaussian Mixture Model (GMM) that best represent the observed data. We then use this model to extract spatial signatures in the collected data, which ultimately enables us to mine transmitter characteristics.

8.4.1 From raw PSD to a GMM

Identifying the best parameters for the Gaussian Mixture Model is not a trivial procedure, particularly in applications such as RF signal analysis, where the input data is inherently noisy. In such cases Expectation-Maximization procedure may suffer over-fitting; that is, more Gaussians are being fitted that there are sub-populations, which results in inaccuracies in transmitter detection. One way to combat over-fitting is by supplying domain knowledge to the fit. In our case, this could be knowledge about the noise floor or the approximate power levels at which we would expect transmitters to appear. While our problem setup does not allow prior knowledge of such information, we can empirically extract the noise floor and approximate transmission powers by employing the following procedures.

Noise floor. A state of the art approach to noise floor extraction is to consider the power level in the guard bands between channels [42]. We take this into account and calculate for each frequency range of interest, what percentage of the range is constituted by guard bands. We then take the corresponding percentile from all the measured data in the band of interest and set this as the noise floor. For example, the noise floor percentile for FM is 4%, while that for TV-UHF is 6.66%. As informed GMM requires both the mean and the variance of a guessed component, we calculate the deviation of the noise as the standard deviation of all values lower than the corresponding percentile. We find that the mean and standard deviation of noise in TV-UHF are -117 and 4.31 and those in FM -90 and 3.76.

Anticipated transmissions. To get a sense of the expected occupancy states in a given frequency band, we perform one non-informed GMM fit. We then consider all the Gaussian components from this fit whose mean is greater than the noise floor. We use the means and standard deviations of those Gaussians due transmissions to inform subsequent GMM fits towards finding the best GMM fit.

Best GMM representation. Now that we have a sense of anticipated noise floor and transmissions, we leverage this information to find the best GMM fit that represents our original PSD data. We do this by fitting a GMM with prior knowledge about the noise floor and anticipated transmitters.

Once we have the best GMM representation of the data we proceed to extracting transmitter characteristics. This is a two-step process in which we first extract transmitter signatures and then analyse these signatures to determine transmitter properties. We describe these two steps in turn.

8.4.2 Extracting signatures

Algorithm 2 Signature extraction.

Require: PSD; $model$
Ensure: Signature
1: **for all** $f \in F, t \in T, g \in G$ **do**
2: $PDF_g(PSD_t^f) = \frac{1}{\sigma_g\sqrt{2\pi}} e^{-\frac{(PSD_t^f - \mu_g)^2}{2\sigma_g^2}}$
{Find association probabilities $AP(PSD_t^f)$.}
3: **for all** $f \in F, t \in T, g \in G$ **do**
4: $AP_g(PSD_t^f) = \frac{PDF_g(PSD_t^f)}{\sum_{g=1}^{G}(PDF_g(PSD_t^f))}$
{Smooth by Belief Propagation (BP).}
5: $SmoothAP = SmoothBP(AP)$
{Extract signatures}
6: $Signature = signature(SmoothAP)$
7: **return** Signature

The goal of signature extraction is to compact the 2D data into 1D space (from ether frequency or time point of view) while using the GMM properties we have obtained. This compact version of the data then allows us to mine characteristics in frequency or time domain. In order to achieve this we employ the algorithm outlined in Algorithm 2, which takes as an input the original PSD data (over F frequency samples and T time samples) and the fitted model (that consists of G Gaussian components), and outputs

a signature comprised of the prevalence of each component for each frequency (or time) bin. The algorithm starts by calculating the PDF of each PSD data point with each extracted component (line 2). It then normalizes these PDFs in order to find the association probabilities $AP_g(PSD_t^f)$ of each PSD value with each component (line 4). Next the algorithm applies belief propagation on each AP_g matrix in order to smooth the effects of noisy data. Finally, we obtain the signatures of each component by calculating for each frequency (or time) bin the prevalence of a component P_g^f as follows:

$$P_g^f = \frac{\sum_{t=1}^{T} AP_g(PSD_t^f)}{T} \tag{8.2}$$

Once we have extracted these signatures we proceed to mining transmitter characteristics. We now describe our approach in frequency domain.

8.4.3 Mining transmitters in frequency

One final step is necessary before we can proceed to extracting transmitter characteristics. That is, we need to make sure that there are not multiple components in the GMM that correspond to the same transmitter. If the latter is true, we need to group these components. To perform this grouping we make use of the calculated signatures and bundle together components that exhibit correlated behaviour. We judge about such correlations based on the variance of each signature across the entire frequency range as well as the power level (i.e. mean) of the component that defines each signature. Particularly, we expect that a transmitter is active in the entire bandwidth of operation, thus we would flag non-contiguous chunks as separate transmitters. Furthermore, if two components are due to the same transmitter we would see correlations between their signatures such as if one signature dips the other one would rise.

We harness these observation and perform a pairwise comparison between all extracted signatures. For each pair of signatures c and cc we check if the means of the two components defining these signatures are within a given PSD threshold from one another. If this is satisfied, we then check for correlations between the signatures. Specifically, we find the standard deviation of each signature (σ) across the entire frequency range as well as the variance of the sum of the two signatures (σ_{c+cc}). We then calculate a score that tells us if the two signatures are correlated as follows:

$$score_{c,cc} = \frac{\sigma_{c+cc}}{max(\sigma_c, \sigma_{cc})} \tag{8.3}$$

If the two signatures are correlated, the score would be at most 1, thus we group components whose score is smaller or equal to 1. The smaller the score, the stronger the correlation between the two components. Finally, in case a component exists in more than one group, we take the union of these groups.

Once we have performed the grouping we can proceed to extracting the number of transmitters as well as their bandwidths and measured power levels. We extract the number of transmitters as the number of continuous occupancy chunks for each signature group. For each transmitter we also calculate its bandwidth based on the frequency spans with non-zero prevalence of the corresponding signatures.

8.4.4 Handling mobile transmitters

When the transmitters are mobile, the temporal variation of PSD values can be large because of the signal attenuation as a result of mobility in addition to the multipath fading and shadowing. Therefore, the cumulative distribution function (CDF) shows a large variation in PSD values of the order of magnitude of 30-50 dB. We use this large variation to flag that transmitter could be a mobile transmitter and further investigate the

temporal variation of PSD values over time to understand the mobility pattern and/or the number of mobile transmitters.

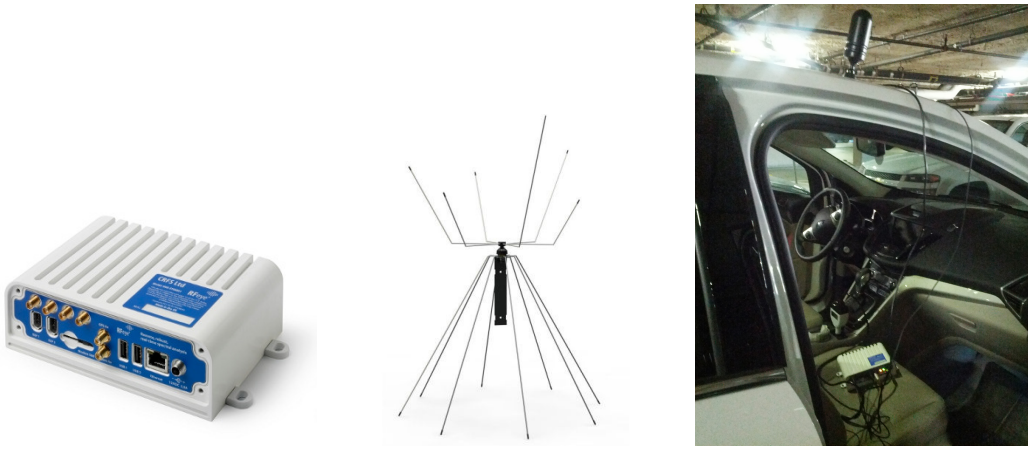
8.5 Evaluation

In this section we evaluate TxMiner. We start by describing the datasets we utilize in our evaluation. We then provide a micro-benchmark to give an intuition of how TxMiner works. We begin our evaluation by looking at accuracy in occupancy detection. Following, we evaluate TxMiner's ability to extract individual transmitter characteristics focusing on transmitter count and bandwidth. We compare TxMiner with a state-of-the-art algorithm for occupancy detection called edge detection [150]. Our evaluation shows that TxMiner outperforms edge detection in both controlled settings as well as in real world measurements. We show that TxMiner has high accuracy in detecting occupancy and individual transmitters' bandwidths. Furthermore, TxMiner is capable of detecting transmitter count and bandwidth in multi-transmitter scenarios.

8.5.1 Measurement setup and data

We collect the measurements required for this analysis using an RfEye spectrum scanner manufactured by CRFS[†] (Figure 8.5(a)). The sensor was configured to scan the spectrum from 30MHz to 6GHz every 3 seconds with variable frequency resolution, depending on the band. For stationary measurements the sensor is equipped with a multi-polarized receiver antenna that supports the entire band from 25MHz to 6GHz (Figure 8.5(b)). For mobile traces we again used the RfEye sensor but this time configured to scan only the TV UHF band and equipped with a magnet-base mobile UHF

[†]`http://www.crfs.com/products/rf-sensor-rfeye-node/`



(a) RfEye sensor. (b) Stationary antenna. (c) Mobile setup.

Figure 8.5: Traces collection was performed using an RfEye spectrum sensor and different antennas depending on the scenario.

antenna as shown in Figure 8.5(c). In addition, we had a GPS antenna connected to the RfEye to be able to sample location along with the measured signal.

We make use of several datasets to evaluate TxMiner. We describe these in turn.

Ground truth. We use measurements in the TV-UHF band as our ground truth. We establish our TV-UHF ground truth by combining spectrum measurements in this band with information from the FCC CDBS, AntennaWeb, TVFool, and white space databases from Spectrum Bridge and iConnectiv, which indicate which channels are idle and which are occupied. The TV-UHF traces were collected using a stationary RfEye sensor scanning the spectrum every 3 seconds with a frequency step of 160kHz. We make use of this dataset in evaluation of characteristics such as accuracy in occupancy detection and detection of transmitter bandwidth. We also use this data to create an artificially mixed dataset (see below) that provides basis for more elaborate evaluation of transmitter count and bandwidth.

Controlled. We utilize a few controlled transmissions to evaluate TxMiner's ability to detect custom transmitters. Particularly, we take traces from three modes of WiMax

transmission: one using 1.75MHz bandwidth, a second using 3.5MHz and a third transmitting at 7MHz. The traces were collected with a stationary RfEye setup scanning every 3 seconds with a frequency step of 160kHz. We also used FCC-certified white space radios from 6Harmonics (Wi-Fi like transmissions in 5 MHz bandwidth) and Adaptrum (proprietary MAC and PHY in 5.5 and 10 MHz bandwidth) with antennas on the rooftop, and used spectrum measurements as ground truth.

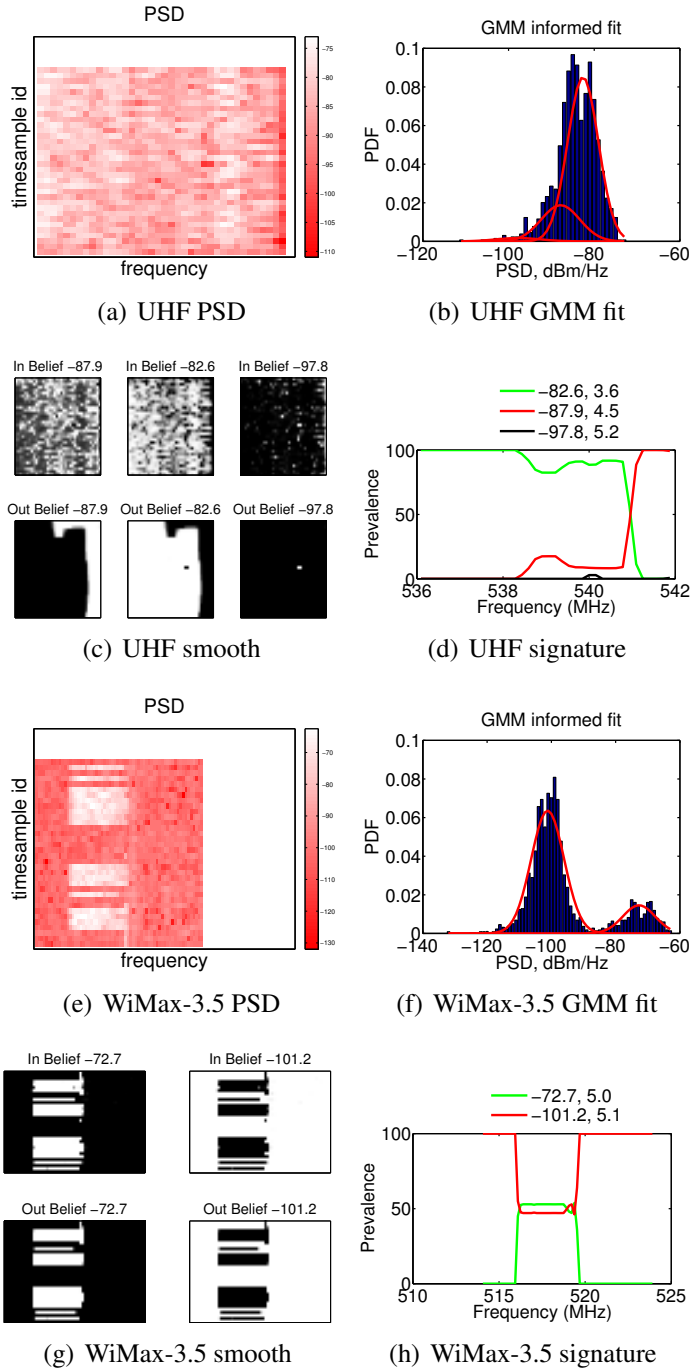
Artificially mixed. We generate artificially-mixed signals drawn from our TV-UHF ground truth. We intertwine over the same frequency band different transmissions or alternate transmission with idle period. By doing so we can emulate single- or multiple-transmitter TDMA schemes, which allows us to establish a ground truth set and quantitatively evaluate TxMiner's ability to tease out multiple transmitters and their bandwidths. Naturally, such ground truth dataset is hard to achieve in a large-scale realistic wireless setup, however, it is not an uncommon scenario.

FM. We also use a dataset in the FM band that was collected using a stationary RfEye scanning the FM spetrum every 3 seconds with a frequency resolution of 40kHz. We use this dataset in evaluation of bandwidth detection.

Mobile setup. Finally, we use a dataset that captures mobility. To emulate a mobile scenario, we used the setup depicted in Figure 8.5(c) to move the receiver to and away from stationary TV-UHF transmitters. We target three transmitters operating on channels 31, 38 and 39. While driving, we scanned the entire TV-UHF spectrum every 3 seconds with a frequency resolution of 160kHz.

8.5.2 Micro-benchmarks

In this section we provide intuition as to how TxMiner works. We analysed many scenarios but in interest of space we present two of them and discuss the outcomes of



(a) UHF PSD (b) UHF GMM fit

(c) UHF smooth (d) UHF signature

(e) WiMax-3.5 PSD (f) WiMax-3.5 GMM fit

(g) WiMax-3.5 smooth (h) WiMax-3.5 signature

Figure 8.6: Micro-benchmark presenting the different stages of TxMiner. The importance of belief propagation in "salt-and-pepper" signals such as the TV-UHF transmission is well emphasized.

each step of the algorithm in each of these scenarios. Particularly we focus on (i) one occupied TV broadcast channel and (ii) a 3.5MHz WiMax transmission.

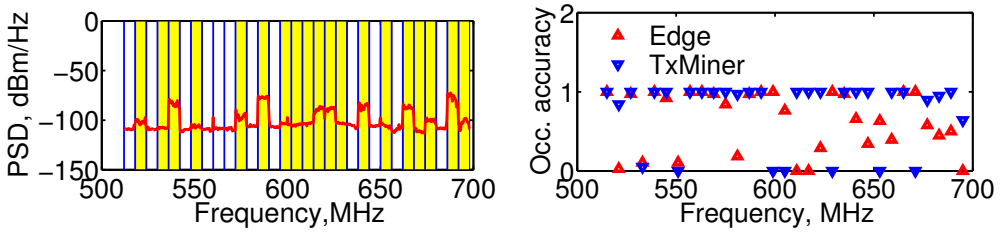
The different stages of TxMiner in each scenario are shown in Figure 8.6. The (a)-figures represent the original PSD data over frequency and time. The more white the color is, the higher the measured power. The (b)-figures present a histogram of the PSD data along with the Gaussian components fitted onto the data. The (c)-figures present results from belief propagation smoothing. The "In Beliefs" are the association probabilities of PSD data with each component (Algorithm 2, line 4), while the "Out Beliefs" are the resulting smoothed association probabilities. Darker colors represent lower values. Finally, the (d)-figures present the extracted occupancy signatures across the frequency band of interest. The legend on those graphs represents the mean and the standard deviation of each of the components.

Our first scenario looks at a band of 6MHz that contains a single UHF transmission. As we can see, in cases such as the UHF transmission, where the channel may be experiencing signal variations across time or frequency, GMM fits multiple Gaussians (in this case three) despite the fact that there is only one operational transmitter (Figure 8.6(b)). Taking a closer look at Figure 8.6(c) we can see that while most of the signal is associated with the $-82.4dBm$ component, there is a persistent signal degradation in the upper edge of the channel that spreads over a larger frequency chunk towards the end of the observation period. This is also reflected in the signature graph (Figure 8.6(d)), where the $-82.4dBm$ component is most prevalent up to 541MHz and then the prevalence switches to the $-88.7dBm$ component. The importance of belief propagation smoothing is also well emphasized in this example: as we can see for all the components the normalized probability matrices are very noisy. Applying belief propagation helps reduce the noisiness and leads to clean signatures, which better allow us to extract transmitter characteristics.

In the second scenario we consider a 10MHz band, 3.5MHz of which is occupied by a WiMax transmission. The presence of both noise as well as transmission PSD values is captured in the GMM fit (Figure 8.6(f)), which depicts two distinct Gaussian components – one centered around $-101.2dBm$ and one around $-72.7dBm$. Since the captured transmission signal is very clear and strong in this case, the GMM fit is much clearer (one component corresponding to each state) and the smoothing procedure does not benefit the algorithm as much as in the UHF case. Finally, as presented in Figure 8.6(h), the signatures of this scenario clearly capture the 3.5MHz transmission and indicate that the rest is idle.

8.5.3 Occupancy accuracy

We begin our evaluation by analysing accuracy in detecting occupancy status. For this experiment we run TxMiner over the entire TV-UHF band (512-698MHz) in 6MHz steps and calculate the accuracy of occupancy detection. In each 6MHz bin there are F frequency samples, depending on the scan configuration. For each of these samples we find if it is occupied or idle. Our accuracy metric then captures the fraction of correctly-detected frequency samples divided by the total number of samples F. Intuitively, we would like our accuracy metric to be 1, since this corresponds to correct detection of all frequency samples in a given bin. For some cases, however, our measurements do not agree with the ground truth established through Google's Spectrum Database and TVFool. Particularly, we measure 7 of the 31 channels in TV-UHF as idle, where they are supposed to be occupied according to the ground truth. In such cases, our accuracy metric would be 0, however this is still a good indicator that TxMiner can persistently detect the occupancy status. In a nutshell, for good accuracy prediction, we want our occupancy accuracy to be either 0 or 1; anything in between indicates weak prediction.



(a) UHF spectrum

(b) UHF occupancy accuracy

Figure 8.7: Occupancy accuracy. We see that TxMiner outperforms edge detection in nearly 50% of the cases.

Figure 8.7 presents our results for occupancy accuracy for the UHF band. Figure 8.7(a) plots the measured occupancy as an average PSD over the capture period. Channels that are supposed to be occupied are designated with yellow. Figure 8.7(b) presents our occupancy results, where the blue markers correspond to TxMiner and the red ones represent edge detection. As we can see, TxMiner typically has a prediction accuracy of 0 or 1 and outperforms Edge Detection in nearly 50% of the cases. For example, channel 22 (the second channel in Figure 8.7(b)) is low-power, thus edge detection fails to recognize it, while TxMiner detects it at 84%. The reason for the poor performance of edge detection is that it often fails to recognize a rising or falling edge, which forces longer frequency spans to be inadequately recognized as idle or occupied.

We use a receiver operating characteristic (ROC) curve to analyze the performance of the transmitter detection method. A ROC curve is a graphical plot which illustrates the performance of a general detection system as some of the parameters (e.g. threshold in our case) is varied. The ROC curve is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate). A ROC analysis gives a holistic view about the sensitivity of the detection method and highlights the true identifyablitly of the algorithm.

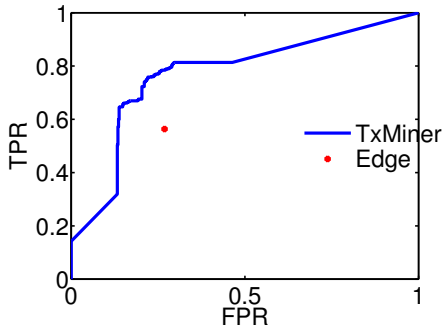


Figure 8.8: ROC analysis of occupancy accuracy. The area under TxMiner's ROC curve is 0.77, which indicates that TxMiner performs significantly better than random. It also outperforms edge detection.

Figure 8.8 presents our results. The blue curve corresponds to TxMiner, while the red mark presents a result for edge detection with 8dBm edge threshold. The area under the TxMiner curve is 0.77, which indicates that TxMiner has significantly better performance than random and clearly outperforms edge detection.

8.5.4 Bandwidth detection

Next we evaluate TxMiner's ability to detect transmitters' bandwidths. First, we focus on our TV-UHF data where we run an experiment in the entire band from 512 to 698MHz in 6MHz steps. At each step we calculate the bandwidth of the detected transmitter. Figure 8.9 presents a comparison between TxMiner and edge detection. The y-axis on the graph presents the ratio between detected and expected bandwidth, where expected bandwidth in this case is 6MHz. As we can see, TxMiner successfully detects the bandwidth of active transmissions and detects a bandwidth of 0MHz where we have measured no transmission or where there is no expected transmission. At the same time edge detection often fails to detect the bandwidth of active transmitters, or detects a 6MHz transmitter in channels that are not occupied. The reason for the poor performance of edge detection is that it often times fails to account for a rising or

falling edge. The latter results in larger areas being detected as idle or occupied than there actually exist.

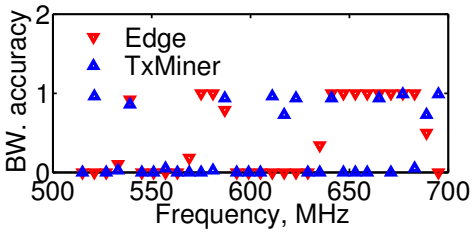


Figure 8.9: Bandwidth detection accuracy. TxMiner is much more accurate than edge detection and successfully identifies the bandwidth of active transmitters. In contrast, edge detection fails in many of the cases by either detecting bandwidth where there is no active transmitter or not detecting anything where there is an active transmitter.

Next we evaluate TxMiner's capability to persistently detect transmitter bandwidth in different transmission scenarios. Particularly we look at the TV-UHF band as well as three WiMax transmissions with known bandwidths of 1.75MHz, 3.5MHz and 7MHz. For the TV-UHF we present average and standard deviation of detected bandwidth across all the channels we identify as occupied. For all the WiMax transmissions we present average and standard deviation across five distinct 100 seconds periods from the captured traces. The expected bandwidths for these transmissions are 6MHz for TV-UHF, 1.75MHz for WiMax-1.75, 3.5MHz for WiMax-3.5 and 7MHz for WiMax-7.

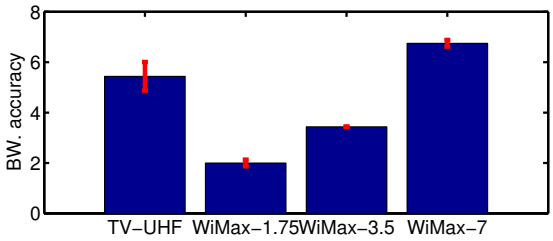


Figure 8.10: Bandwidth detection across different transmitters. TxMiner is persistently able to detect the bandwidth of different transmitter and the detected values are very close to the expected ones.

Figure 8.10 presents our results. As we can see, TxMiner is persistently able to recognise the bandwidth of each transmitter type. Furthermore, the detected bandwidths are very close to the expected bandwidths.

8.5.5 Detection of multiple transmitters

Next we evaluate TxMiner's performance in scenarios where multiple transmitters are present. To emulate such scenarios we artificially mix and amplify measured signals.

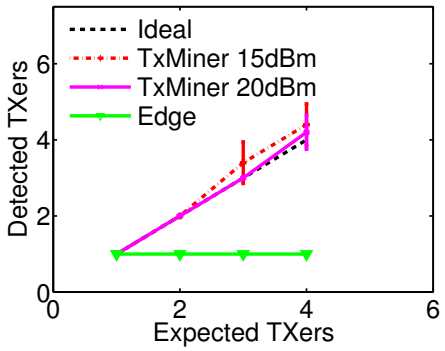
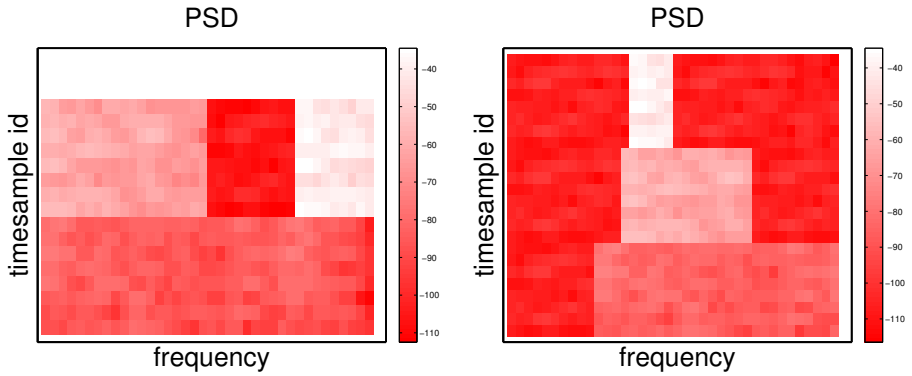


Figure 8.11: Transmitter detection with increasing number of transmitters. TxMiner is able to detect the number of transmitters as they increase and clearly outperforms edge detection, which cannot identify more than one transmitter.

Our first evaluation focuses on TxMiner's ability to detect an increasing number of transmitters of the same bandwidth. For this experiment we mix measured signals from the TV-UHF band and artificially amplify them (by adding 15 or 20dBm) to make the difference between transmitters more pronounced. We then run TxMiner and count the number of detected transmitters. Figure 8.11 plots the number of detected transmitters as a function of the number of expected transmitters. We present two results for TxMiner: one where the separate transmissions are 20dBm apart (TxMiner 20dBm) and one where the transmissions are 15dBm apart (TxMiner 15dBm). We also com-



(a) Case 1 (b) Case 2

Figure 8.12: Evaluation cases of multiple transmitters with different bandwidths.

pare TxMiner's performance with edge detection. As we can see, TxMiner clearly outperforms edge detection. The reason for the poor performance of edge detection is that it only considers an average of the measured signal and unlike TxMiner, does not take into account the spatial properties of the signal. In contrast, TxMiner is capable of detecting the number of transmitters with high accuracy. We see that the accuracy of TxMiner decreases as the number of expected transmitters grows. This trend is more pronounced in the 15dBm margin scenario, where the GMM fit is less able to differentiate between signals. Thus, in order to assure high accuracy of teasing out number of transmitters, one needs to take smaller time samples at a time and infer about occupancy based on multiple consecutive observations.

Next we evaluate TxMiner's ability to extract multiple transmitters with variable bandwidths. For this experiment too we use artificially mixed and amplified signals. We study two cases of spectrum occupancy presented in Figure 8.12. Each of these cases includes a different configuration of three transmitters. In case 1 we have a 25 second transmission with 6MHz bandwidth, followed by two concurrent transmissions, one 3MHz wide and one 1.4 MHz wide and separated by an idle zone. The second case features three consecutive transmissions each of 25 seconds. Table 8.1 presents

Table 8.1: Detection of multiple transmitters with variable bandwidths.

	TX 1		TX 2		TX 3	
	E.BW (MHz)	D.BW (MHz)	E.BW (MHz)	D.BW (MHz)	E.BW (MHz)	D.BW (MHz)
Case1	6	5.68	3	2.84	1.4	1.26
Case2	4.375	4.26	2.34	2.21	0.78	0.63

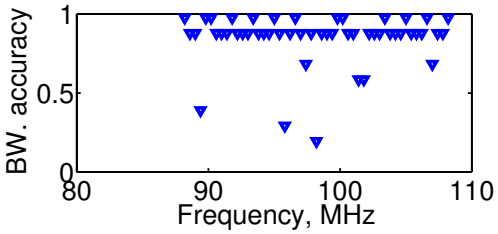


Figure 8.13: Bandwidth detection in the radio FM band.

for each case and each transmitter the expected and the detected bandwidth (E.BW and D.BW, respectively) . As we can see, TxMiner successfully detects all the expected transmissions and is also accurate in detecting their bandwidths.

8.5.6 FM band

We also evaluate TxMiner's performance in cases with narrow band transmissions such as those in the radio FM band. TxMiner detects this entire band as occupied. Figure 8.13 presents our results for accuracy of bandwidth detection. In this experiment we ran TxMiner over the entire FM band from 88MHz to 108MHz in steps of 400kHz. The graph presents for each 400kHz chunk the bandwidth accuracy expressed as the ratio between detected bandwidth and step size (in this case 400kHz). As we can see, majority of the detected channels have bandwidth accuracy of either 0.98 or 0.88, which corresponds to a bandwidth of 392kHz and 352kHz, respectively. The 392kHz bandwidths likely correspond to HD radio transmissions, which by specification occupy wider bands. The 352kHz transmissions correspond to stations that were sensed

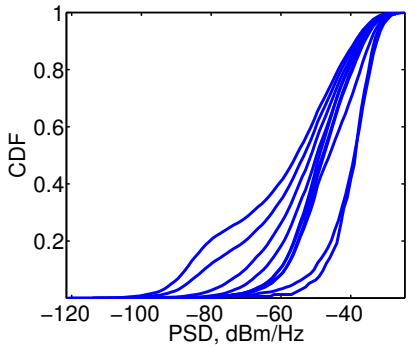


Figure 8.14: Mobility. The shift of the CDF as well as the decrease of the slope indicates a decreasing trend of PSD over time, which signifies mobility in a direction away from the scanned transmitter.

with very strong signal, in which case we would see the squelch tones as a separate peak. Finally, we see transmissions whose bandwidth accuracy is lower. Those are likely to be radio transmissions that were sensed with low power, thus their bandwidth does not span the entire 400kHz band.

8.5.7 Mobility

In this section, we evaluate mobile transmitters using the dataset collected from controlled experiments in mobility. The CDF of the dataset collected over channel 39 in TV-UHF band exhibits a variation in PSD of as much as 60dBm over a span of 30 min. Therefore, we further investigate the temporal variation in CDFs over time. Figure 8.14 shows ten CDFs obtained from PSD values in first 3min, 6min, 9min, and so on up to 30min of PSD data. We observe that the CDFs shift toward the left and exhibit a lower slope indicating a much large variation in PSD values over time. Further, in this particular case, the decrease in means of probabilistic distribution of the PSD values suggests that the receiver is moving away from the TV transmitter in channel 39.

8.6 Related Work

Prior work on spectrum analysis can be classified into 3 categories: wide band spectrum occupancy analysis, envelope detection for identifying unknown signals, and detecting transmitters with known signatures.

Several studies have analysed large scale spectrum measurements to identify portions of spectrum that are not used [99, 145], or identify patterns of primary users such that unused spectrum can be opportunistically reused [39, 87]. This body of work assumes no knowledge about the transmitter. They typically apply a threshold for noise, and any signal above this threshold is assumed to be occupied, anything below is assumed to be free. [39] analyses spectrum from China, and models the arrival of users in the cellular bands. [87] analyses spectrum from 30 MHz to 6 GHz, and studies opportunities for dynamic spectrum access in these bands. However, none of these analyses share the goal of TxMiner, and are unable to predict transmitter characteristics from a wideband spectrum trace.

Another set of techniques, which is primarily used by practitioner, is to tease apart unknown transmissions from known transmitters. This is frequently used to identify interferers in the spectrum, for example, in the wireless carrier spectrum. The most common technique is that of envelope detection. A circuit (or these days software) tries to fit a curve around the max-hold (or mean) of signals. Although this technique is useful in determining anomalies in the curves, it does not provide much insight into the distributions that make up the max-hold or mean.

The most closely related work to TxMiner is DoF [68] and AirShark [125]. DoF builds cyclostationary signatures for different transmitters in 2.4 GHz, such as Wi-Fi, Bluetooth, etc., and mines spectrum data for these signatures to determine the users of the spectrum. AirShark tried to solve a similar problem, but using commodity Wi-Fi chipsets. While both these techniques are useful, they only work when the transmitter

patterns are known. TxMiner takes the next step, and identifies transmitters when their patterns are not known.

8.7 Discussion & Future Work

Although the knowledge gleaned by TxMiner is very useful, it is still the first step. We believe that many more details can be learnt about transmitters, which will enable several additional applications of spectrum analysis. We list some of our research efforts in this direction below.

Collocated transmitters: Since TxMiner looks at power profiles of transmitters, it is unable to distinguish two collocated transmitters operating at the same power level, but on adjacent bandwidths. Both those transmitters will be classified as a single wideband transmitter. This commonly happens for FM broadcasts. We currently use knowledge of current occupants, e.g. FM radio broadcast, to determine these transmitters, but a more automated approach is still an open problem.

Mobile transmitters: Although TxMiner is able to infer whether there is a mobile transmitter, its accuracy in determining the number of mobile transmitters is limited. This is because a mobile transmitter can be confused with many static, and mostly inactive, transmitters. Another open question is the speed of the mobile transmitters. We are actively investigating techniques that break the spectrum trace into smaller chunks, and might help us answer these questions.

Integration with Known Transmitter Signatures: As mentioned in the previous section, prior work [68,125] has looked at identifying transmitters with known temporal signatures. We are actively investigating approaches to leverage the prior work for a more accurate spectrum analysis system. In particular, one approach we are studying

is to perform DoF on the time samples before the FFT, and only perform FFT on time samples that do not have a known temporal signature.

Determined attackers: We note that a determined attacker can try to beat TxMiner by using a different power level for every packet.[‡] However, we note that the attacker is limited by the number of power levels at his/her disposal. A very high transmit power will make it easily detectable using simple observation, and a low power level might be insufficient to establish a communication link. Given the limited number of transmit power levels, TxMiner will identify the same transmitter as a small number of unique transmitters. Bundling these together into one transmitter is part of future work.

To summarize, in this paper we have presented a system, called TxMiner, that is able to mine raw spectrum measurement data and identify properties of transmitters operating in that spectrum. TxMiner is based on a simple observation from Physics that fading is log-normal. We use this principle to build machine learning algorithms (GMM with Belief Propagation) that can attribute spectrum measurement points to the different transmitters. Our results show high accuracy in detection of transmitter count and bandwidth as well as mobility patterns. To the best of our knowledge, TxMiner is the first system that can achieve this goal.

8.8 Acknowledgements

The material in this chapter is based on joint work with Aakanksha Chowdhery, Ranveer Chandra, Ashish Kapoor, Paul Garnett, Paul Mitchell and Elizabeth Belding.

[‡]We note that existing communication systems, including CDMA, do not use a different power level for every packet.

Chapter 9

Applications: ImmuNet – Improved Immunization Through Cellular Network Technology

9.1 Introduction

Vaccinations are one of the greatest public health achievements of the 20th century. They not only reduce the rate of death and disability due to preventable diseases but also improve the productivity of workforces. While vaccination has lead to the successful eradication of a number of diseases in the developed world, developing countries still fall behind in implementing high-coverage immunization strategies. According to statistics from the World Health Organization*, 1.5 million children died in 2010 from diseases preventable by vaccines. The highest percentage, 42%, of the cases were in the African region and 29% were in South East Asia. Most of the deaths were caused by Hib Meningitis, Pertussis, Measles, Pneumonia, Tetanus and Rota-virus; all of these are diseases that can be prevented through vaccination. Furthermore, in many cases vaccines for these diseases are available; however, vaccination schedules are difficult to

*http://www.who.int/immunization_monitoring/data

enforce due to missing or incomplete information about individuals who need immunization.

There is increased effort in incorporating Information and Communication Technology (ICT) to improve vaccination distribution to children and infants [41, 81, 113]; however, such systems are often designed according to western models and operate under assumptions that do not hold in developing countries. Vaccination practices often differ from country to country and between urban and rural areas in developing regions. For example, our preliminary survey with health workers from the rural village of Macha in Zambia, reveals that, due to a limited work force, there is no tracking of individual immunization status currently in place. Instead, vaccination targeting and bookkeeping is done at the community level. To facilitate efficient distribution of vaccinations, there is a need for a system that is self-sustainable, can operate without health worker intervention and can function as an addition to day-to-day vaccination routines.

In order to facilitate such efficient distribution of vaccinations, we propose ImmuNet, a system that leverages cellular network technology and database tracking to keep individual immunization records. As most mandatory vaccinations occur between the age of 0 and 5 years, ImmuNet is designed for improved distribution of vaccinations to infants and children under the age of five. This focus, coupled with the specific vaccination practices in developing countries, poses unique challenges in the system design. To operate in rural areas with low availability of personal records, the system enables collection of personal data in digital format that is submitted either by health workers or directly by the patients. By utilizing local cellular network infrastructure, the system provides prompt dissemination of information for vaccine availability and schedules in the form of text messages. ImmuNet can also record users' network association time and location and thus build patient interaction graphs to predict disease spread patterns and vaccination needs. Based on these social graphs, ImmuNet provides an alert system

in the form of a heat-map that notifies health workers for high risk regions that need vaccinations.

The majority of the population in sub-Saharan Africa lives in rural areas. People who live in rural areas typically have to travel much longer distances in order to reach a health care facility [56, 110]. Hence, a system like ImmuNet that provides notification of vaccine needs and availability would be extremely beneficial in prompting people from rural communities to get their children vaccinated and would enable people to synchronize their travel to the closest health care facility offering needed vaccination. Following these observations, our design specifically targets remote rural areas in the developing world. At the same time, our system is flexible and can be successfully applied in urban areas with different levels of availability of individual records.

To build an end-to-end proof of concept system, we leverage *VillageCell* [20], a low-cost cellular network architecture that provides local voice and text services for free. ImmuNet, however, is not restricted to operating only with VillageCell; instead ImmuNet is a generic system and can be integrated with any cellular network provider. Our system provides a text message interface for the collection of immunization information directly from the patients and enables quick immunization status verification and targeted dissemination of vaccines. By tracking the migration of cellphones in the network the system can infer human interactions and consequently possible disease spread patterns in rural regions. For data storage, ImmuNet utilizes a database system, called *VaccStore*, that collects patients mobility information as well as vaccination and demographic information. A data analysis engine on top of VaccStore detects individuals to be targeted for vaccination and notifies them via text messages. Where individual vaccination records are not available, initial patient information must be collected to bootstrap ImmuNet. As ImmuNet is designed for improved vaccination distribution specifically to children and infants, a crucial part of the information supplied by pa-

tients associated with the system is their family relational data. This data, coupled with the biometric data collected for identifying individuals, is central to the operation of *VaccStore*.

9.2 System overview

ImmuNet is designed to provide technological support related to vaccination to both health workers and patients. Through ImmuNet, health workers can collect patient data directly in digital format, a feature that is widely incorporated in recent health care technologies [46,48,90,113] for reduced paper work and higher flexibility in accessing and processing patient data. The data collected through ImmuNet will enable a plethora of features for the benefit of health workers, such as quick verification of immunization status and a global overview of the vaccination needs in a geographical area in the format of disease spread *heat-maps*. Furthermore, ImmuNet is designed to enable health workers to send notifications as unicast or broadcast text messages to patients using existing cellphones, so individuals or groups from the community can be targeted.

In remote rural areas, however, the health workforce is often not enough to collect and maintain personal vaccination records. In such cases, vaccination tracking is typically done at the community level. To enable ImmuNet's operation in such environments we develop functionality that allows data collection directly from individuals through text messages. This data is then stored in the database and processed to determine individual vaccination needs. The system then automatically sends text messages to persons or groups who are in need of available vaccines. The independence of ImmuNet in collecting data and sending notifications makes it suitable for operation even in areas with low availability of personal immunization records. Furthermore, health

workers can still benefit from quick lookups of immunization status and heat-maps, which can be done based on the information collected directly from patients.

Along with the aforementioned features, we identify a set of goals that our design achieves, to establish ImmuNet as a system that is widely suitable across different countries in the developing world. First, ImmuNet is a low cost system that can run on a single server (assuming that the cellular network is already in place) and uses only free and open source software. The system is easy to implement and self-sustainable. It can operate with any cellular network without modifications of users' handsets. Through social surveys and user feedback, our goal is to design ImmuNet to be well aligned with the needs and abilities of people from the community. As it operates with personal data, ImmuNet should address security and privacy issues; hence, special effort should be put in deploying techniques for enhanced security in handling data. In this paper, however, we focus on the initial system setup, leaving security considerations as a future direction of our work.

9.3 Related Work

There are a number of projects that leverage network technology for enhanced health care. The areas of application vary from telemedicine [97, 136], approaches to increase health workers motivation and efficiency [47, 48, 122], patient data collection [41, 46, 48, 81, 90, 113] and tracking of vaccination distribution [114].

From the large body of work that leverages technology for improved health care, the closest to ours are [81] and [41]. In [81] Klungsøyr et al. propose a system called mVAC that is based on openXdata [113] and implement an end-to-end mobile phone-based solution for recording individual immunization information in digital format. The ChildCount++ [41] project also develops a cellular network based system in which

health workers utilize cellphones to record individual patient data and a custom computer program to digitalize data from old paper forms. These systems are different than ImmuNet. mVAC, for instance, requires each user to have a special card with 2D barcode for identification. In contrast, ImmuNet does not require users to possess any additional items for identity verification; instead it uses cellphone unique ID for user identification.

9.4 Research challenges

There are a number of challenges in the design and deployment of a system for ICT support of vaccinations in remote rural areas. In this section we present the main challenges, which we divide into two categories: *technical* and *anthropological*.

9.4.1 Anthropological challenges

In remote rural areas the population density is often very low. A village is typically spread over a large geographic region and the population lives in small clusters, a few kilometers apart from each other. This low population density poses a challenge in building networks to connect individuals in remote rural areas. Furthermore, it brings low incentive to commercial cellular providers to deploy networks in such areas, which results in sparse cell phone coverage.

The low income of people in rural communities is yet another important factor in the design of a technology for improved distribution of vaccinations. In Africa, about 70% of the population lives with less than $2 per day[†]. At the same time, prices for mobile cellular connectivity are five times higher than those in the developed world[‡]

[†] `http://data.worldbank.org/`

[‡] `http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf`

and, while mobile cellular services are being rapidly adopted in developing nations, the personal expenses related to this adoption still constitute a significant portion of the family budget.

Another important factor is the likelihood of people to approve and adopt a system that utilizes technology for improved vaccination distribution. The percentage of population in the developing world who own a cellphone increased dramatically over the last five years – from 32% to 79%[§]. Data indicates high adoption of cellular technology and also of text messaging due to its cost efficiency. Hence, cellular network technology has excellent potential as a platform for development of health care support systems in developing regions. Regardless of the intermittent cell phone coverage, the benefits of cellular technology are well understood and people are often ready to travel to an area with coverage in order to communicate [45].

9.4.2 Technical challenges

ImmuNet consists of two primary inter-operating components – a *database* and a *cellular network*. The database stores patient information. A data processing engine periodically scans this database and generates immunization alerts and notifications to be sent to patients. Through the cellular network, ImmuNet collects patient information and distributes messages related to immunization. The deployment and operation of these two components in rural areas, however, poses some unique challenges, which we discuss next.

The sparsity of population, coupled with the low income of people in remote rural areas, brings little incentive for commercial providers of telecommunication services to build networks in such areas. As a result, commercial cellular coverage is often not

[§]http://www.itu.int/ITU-D/ict/statistics/

available. Furthermore, commercial cellular services such as text messaging are not free.

To tackle these challenges and build an end-to-end proof of concept system, we leverage VillageCell [20], a low-cost local cellular network architecture. VillageCell utilizes low cost radio hardware and free open source software to provide local voice and messaging services. ImmuNet, however, is not restricted to operating only with VillageCell; instead, ImmuNet is a generic system that can be integrated with any commercial cellular provider.

In remote rural areas in Africa, individual immunization records are often not kept. Thus, we cannot use data from already existing records to bootstrap ImmuNet. To overcome this challenge we plan to incorporate text message based functionality in ImmuNet to prompt patients to submit their (and their childrens') personal identification and immunization data the first time they connect to VillageCell. ImmuNet then extracts this data from the submitted text messages and populates its database called VaccStore. As a result, the system will be able to operate without the reliance on health workers to populate the database with patients records. At the same time, however, relying on patients to submit their personal information has one major drawback. That is, patients may submit incomplete or incorrect information which would lead to populating VaccStore with uncertain and noisy data.

The problem of uncertain or incomplete data has been a focus of interest in the database research community in recent years and there are several approaches proposed under the umbrella of probabilistic database theory [24, 36, 44, 135]. In *VaccStore*, we implement a relational database that can capture uncertain data in an efficient manner by implementing arbitrarily complex queries while keeping the data schema simple. This approach is similar to [135] and allows for a flexible framework where the system can easily adapt as we gain more knowledge about which pieces of data are easy or hard

to attain in the context of immunization in rural areas and therefore should be modeled as either deterministic or probabilistic.

9.5 Architecture

To overcome the problem of sparse cellular coverage and build a proof of concept ImmuNet system, we utilize *VillageCell* [20], a low cost GSM network that provides local call and text messaging services for free, which is critical to impoverished residents and those with unpredictable income. Furthermore, the system allows users to utilize their existing handsets without any modification, which is extremely important for the widespread adoption of the technology. We implement a database dubbed *VaccStore* that collects patient information and generates a variety of notifications, reminders and alerts based on the collected mobility and relational data.

An overview of the ImmuNet architecture can be seen in Figure 9.1. We now present more details of the two main components of ImmuNet: VaccStore and VillageCell. Furthermore, we present details about their integration to support ImmuNet.

9.5.1 VaccStore

A key part of ImmuNet is the VaccStore database that stores personal biometric and identification data for each person, as available, and their immunization status. Our goal is to provide a system that is easy to use and efficient. For this purpose, we implement VaccStore in MySQL, one of the most commonly used relational database management systems. Relational databases provide a declarative method for specification of data and queries. This property, coupled with dynamic views and stored procedures provided by relational databases, provides a means for a dynamic and flexible framework. Through a MySQL implementation, VaccStore also supports various users and

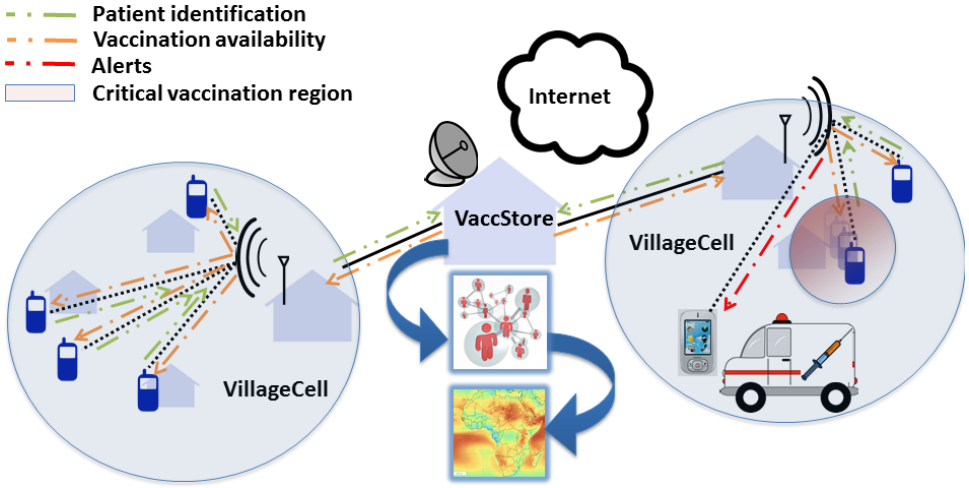


Figure 9.1: ImmuNet Architecture: ImmuNet uses *VillageCell* as a cellular network platform and *VaccStore* as a database to collect patient identification, to distribute immunization information and to generate high-risk alerts.

user rights crucial for a system with strong security. In addition, VaccStore enables efficient data access and updates with the use of indices. However, the database querying imposes novel challenges, as the search has to be performed on a noisy dataset. Such challenges cannot be handled with a traditional relational database solution. Therefore VaccStore extends this traditional paradigm to adapt to the requirements presented in the context of vaccination in developing regions. There are two main ways in which the relational database paradigm needs to be extended: first a probabilistic data model should be introduced to capture incomplete or uncertain data and, secondly, basic diffusion models necessary for capturing the spread of diseases and influence should be incorporated to create a *social* database. Here we describe the main VaccStore design choices and how they address various research challenges.

Data features: The main VaccStore schema consists of six tables that are listed below. The relations between these tables are provided in Figure 9.2. This basic schema will be extended to account for uncertain or missing data as described later.

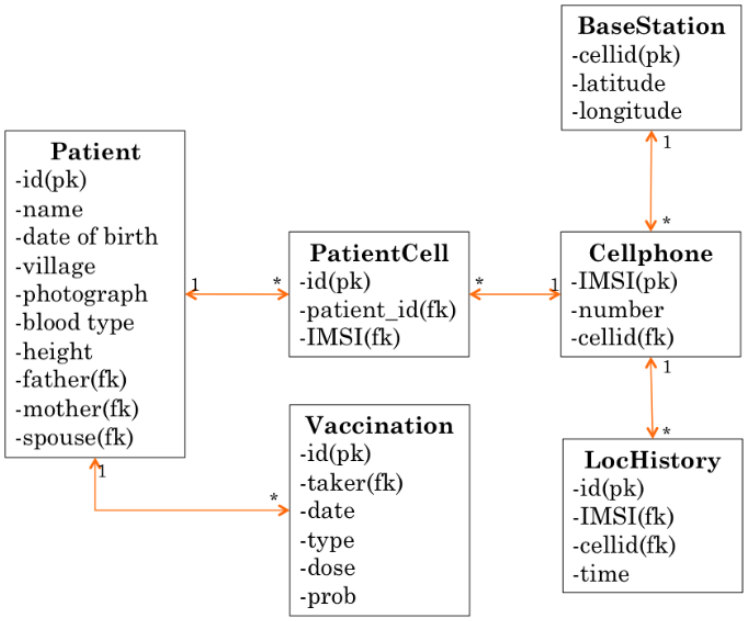


Figure 9.2: VaccStore Table Relations: VaccStore uses six main tables. Here *pk* means primary key while *fk* means foreign key of the given table.

1. Patient: contains contact, identification, family relation, and biometric information for each patient.

2. Vaccination: contains vaccination records for each patient. Given that individual level vaccination status is hard to attain correctly, each vaccination record has a probability associated with it (*prob*) capturing the certainty of such a vaccination.

3. Cellphone: contains mobile phone information which is held by common subscribers/patients.

4. BaseStation: contains VillageCell base station information.

5. LocHistory: contains roaming history of a mobile phone.

6. PatientCell: contains the patient-cellphone ownership mappings. This mapping is of many-to-many nature, meaning a patient can have many cellphones and a cellphone can be shared by many patients.

Incomplete and uncertain data: Due to the scarcity of immunization and identification information in rural Africa, a traditional relational database scheme where every record stored exists with *certainty*, and every field in that record has a precise, unambiguous value, is not applicable. For instance, at a given time a given user being immunized for a given illness or being infected with one, is not known with certainty. Uncertain data has been an active area of research in the database community [24,36,44]. Our goal is to represent the uncertain data by a combination of classical database relations and propositional formulas, a method that is influenced by graphical models [70, 135]. We aim to investigate which database attributes are difficult to collect correctly and completely for the entire population and provide a flexible framework that can handle *attribute-level uncertainty*. Similarly, collecting data about certain members of the community can be harder than others, therefore creating a *tuple-level uncertainty*. For both cases the uncertainty is expressed in terms of probabilities: either a tuple is present only with some probability, or the value of an attribute is given by a probability distribution. Therefore, we implement a simple relational database design that can be extended to incorporate a probabilistic data model where certain columns (or tuples) have probabilistic rather than deterministic values. In this design, the base tables still have a simple model avoiding unwanted dependencies while the queries that involve this uncertain/incomplete data can be quite complex providing a flexible analysis tool. Such a design decision also allows us to build the VaccStore framework on decades of research on relational database management systems even in the face of uncertain data.

Dynamic data: VaccStore stores highly dynamic data such as geo-locations of its users as well as their immunization status. Storage of such data allows the sys-

tem to evaluate geographical risk zones and notify its users accordingly. For instance, VaccStore is designed to frequently communicate with VillageCell to capture current location of users and automatically re-calculate risk factors for regions when the population residing in that particular region changes. Similarly, when a user moves to a region with a high risk factor, VaccStore automatically detects and prepares a warning message that will be sent to the user. Such use cases are implemented through database triggers.

Biometric data: Complete and accurate personal information is sparse in remote regions, especially due to a lack of human work force and technological solutions. This in turn creates a big challenge: "how can a person be uniquely identified without such accurate information?" We use biometric information that is easy to obtain and process for this purpose, namely photographs that can be captured through mobile phones. We use EigenFace [141] as a face recognition algorithm. The algorithm works as follows: given example face images for each person in the database, plus an unknown face image to recognize, it first computes a "distance" between the new image and each of the example faces. Given the set of distances from each of the existing images, it selects the example image that is closest to the new one as the most likely known person. If the distance to that face image is above a threshold, the image is recognized as that person; otherwise, it is classified as an "unknown" person.

We tested the face recognition feature of *VaccStore* using 400 benchmark images (40 different people). The recognition success rate is more than 70% when only one image for each person is fed into its training process. Given the size of the population, the first step of computing the distances for "all" images can get expensive. Therefore, we also associate a set of characteristics for each image such as age range, geographical range. As future work, we plan to use these features to perform filtering based on these predicates in order to provide a more scalable solution.

Building a social database: One of our main goals is to create a database that is a natural fit for processing social data. Health related issues are social in their nature. Identifying diffusion of information or diseases requires building and using algorithms that rely on social relations. Through stored procedures that are built to support queries relating to such problems, we aim to provide good performance in the face of large scale data. Examples of *social* queries include *selecting mothers/brothers/uncles of infants between ages X and Y and are not immunized* or *selecting top k influential people in a given community*.

A comprehensive understanding of *influence* in social networks is vital for an effective deployment of ImmuNet. There are two types of influence that are implemented through stored procedures in VaccStore: *local* and *global* influence. Computation of *local* influence is relatively simple and refers to the total influence a given person has in its immediate neighborhood, for instance immediate relatives. Since a person influencing her immediate neighborhood can trigger further diffusion of influence, a more interesting problem is to define and calculate the influence a given person has in the entire network. In order to capture such global influence, we will provide database functions that use family relations as well as colocation information extracted from the location history table.

Initializing VaccStore: At the onset, the database can be initially seeded through SMS queries to users over the VillageCell architecture to collect identification data and existing vaccination information. In addition, biometric data, such as photographs, and current identification information can be input through smartphones utilized by medical professionals, at the time of vaccination.

9.5.2 VillageCell

VillageCell's network architecture, first described in [20] and summarized here, is presented on Figure 9.3. Each base station consists of two components: *(i) radio hardware* – we use Universal Software Radio Peripheral 2 (USRP2)[¶], which is a commercial Software Defined Radio (SDR) platform, and *(ii) a general purpose PC*. USRP2s can operate in a wide range of frequencies, which allows us to use any of the GSM bands. Each VillageCell base station runs a free open source implementation of the GSM stack called OpenBTS. Calls are routed within or outside the local GSM network through Private Branch Exchange (PBX) servers, which run open source free software called FreeSwitch[‖]. The PBX server also maintains information about user association and mobility patterns. To provide text messaging functionality in the system we utilize advanced OpenBTS modules *smqueue* and *sipauthserve*. The former component provides a store and forward SMS queue, which is capable of storing messages for a certain amount of time before delivering or discarding them. This functionality is extremely important, as we expect that users will typically not be in range at all times. Thus, their messages can be delivered in a delay tolerant fashion whenever they re-associate with VillageCell.

The components of VillageCell are connected over a wireless backhaul, which is essential for a system that operates in rural remote communities with no infrastructure. Currently, our system prototype uses off-the-shelf Wi-Fi devices to establish connections between the VillageCell components. However, we are working on developing a new technology that utilizes the frequency range from 50MHz to 800MHz, referred to as *white spaces*. We chose white spaces because the propagation characteristics of

[¶] http://ettus.com/
[‖] http://www.freeswitch.org/

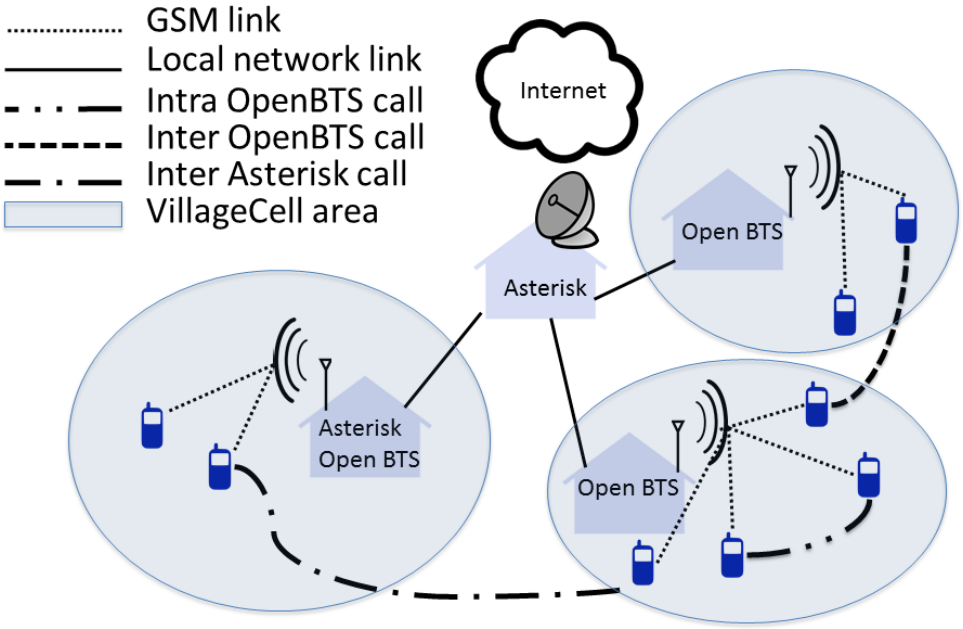


Figure 9.3: VillageCell Architecture: VillageCell's base stations run OpenBTS. Calls are routed in or between base stations through PBX servers running open source software (FreeSwitch in this implementation).

this band allow wide coverage in rural areas whereas traditional Wi-Fi links experience problems due to increased interference and lack of line-of-sight.

VillageCell enables a number of features that are of critical importance for a system for improved vaccination distribution and tracking. First, it allows the collection of patient information directly from the patients themselves. Furthermore, it allows free dissemination of information for vaccine availability and schedules. The ability of the system to record user association time and location allows for building a dynamic social graph of human interactions and predicting disease spread patterns. Furthermore, this functionality allows implementation of an alert system that, on one hand can prompt patients if they are entering a high risk area and, on the other hand can alert health workers of areas with increased health risk, so vaccinations can be administered promptly.

9.5.3 Integration

VillageCell and VaccStore are integrated together in our prototype to enable the end to end operation of ImmuNet. When a user first associates with VillageCell, ImmuNet recognizes the user as new and initiates a process for new user registration. This process includes the exchange of text messages in which the system asks the newly associated user for information related to immunizations and the user can choose whether to answer, and thus, whether to register with ImmuNet.

In VaccStore, there is a daemon that periodically polls VillageCell's PBX server for patient mobility information. Data such as (i) SIM card unique identifier (IMSI), (ii) associated base station ID, (iii) time-stamp and (iv) location (if available) are recorded in a local file and imported in VaccStore's PatientCell and LocHistory tables. In the opposite direction – from VaccStore to the users associated with VillageCell – the system can supply various notifications and alerts in the form of text messages. These text messages can either be sent manually by health workers through a specially designed interface, as described below, or automatically from the system. For example, if a person enters an area with a high risk of a certain disease, the system can automatically notify them based on mobility information. The system can also generate alerts to health workers. For example, a child is at a given age and needs to be immunized, a responsible health worker is notified and she can send an individualized message to the caretaker of the child.

Based on the collected patient data, the system can build a social graph of human interactions and a heat-map of possible disease spread (Figure 9.1). This heat-map is useful in the process of identifying areas with high risk that need prioritized immunization.

For enhanced text based communication we build upon VillageCell's basic text messaging functionality by providing a custom computer application that converts Instant

Messages (IM) to SMS. This application provides a quick and easy way for health workers to type messages and send them to individuals or groups of patients. Furthermore, our IM client exposes an API to VaccStore so VaccStore can generate and send messages automatically upon a trigger event without health worker intervention.

9.6 Current status and future outlook

We have two avenues for experimental testing of our system. The first is in-house at UCSB. Once successful testing is completed at UCSB, we will leverage our partnership with the MachaWorks[**] organization to deploy a small scale test in the rural village of Macha, Zambia. We will use this deployment to perform in-situ testing and integration with existing networks. Through this partnership, we have previously deployed test networks and have studied the performance of existing wireless network deployments [75, 78].

In the second stage of the project we will concentrate on the social behavior tracking and information cascading. We will use our deployment in Macha, Zambia to analyze different schemes for information distribution with respect to the mobility patterns that exist in the area. In our previous work, we devised a method for identifying and limiting the impact of malicious agents in online social networks [32]. We plan to develop, deploy and test similar algorithms that use VaccStore personal and biometric data, correlated with VillageCell point of attachment data, to suggest which individuals and areas should be prioritized during the vaccination process.

For ImmuNet to be realistic and well aligned with user needs and expectations, we plan to conduct a series of social surveys with residents from Macha. We envision three approaches to surveying the local community: (i) on-site interviews, (ii) Online Social Networks (OSN) questionnaires and (iii) SMS surveys. Part of our research team

[**] `http://www.machaworks.org/`

is traveling to the rural village of Macha in June 2012. Along with our deployment work, we will have conversations with people from the community to gain insights into immunization practices, perceptions of immunizations in general and the likelihood of adopting technology for health care. To be able to assess the level of adoption of ImmuNet, without being physically present in the community, we also plan to utilize social surveys through OSNs as well as obtain feedback through free local SMS.

9.7 Conclusion

Vaccinations have helped mitigate or eradicate numerous serious and deadly diseases in the developed world. At the same time, many developing countries still fall behind in implementing efficient vaccination practices. The combined effort of health workers, computer scientists and engineers to incorporate technology for improved health care has had tremendous success over the last few years. ImmuNet carries on in this direction by providing technology, accessible at no additional end user cost, for improved distribution of vaccinations to children and infants in developing regions.

As described in Section 9.4, incomplete information for individual immunization, coupled with the health care personnel shortages, in remote rural areas are formidable problems for the implementation of high coverage immunization strategies. Thus, designing intelligent technologies with minimal requirements for health worker involvement is a promising solution for improved distribution of vaccinations. We believe that with its flexibility in data collection and interaction with patients, ImmuNet has excellent potential for impact on the way vaccinations are distributed in rural areas as well as how vaccination status of patients are tracked. Through our existing partnership with the MachaWorks organization in Zambia we will initiate system deployment in the field in order to emphasize the practical impact of such system. We hope that our findings

will inspire wider deployment and adoption of ImmuNet in the near future, as well as the design of other health care personnel-independent systems.

9.8 Acknowledgements

The material in this chapter is based on joint work with Ceren Budak, Arghyadip Paul, Biyang Liu, Elizabeth M. Belding, and Amr El Abbadi.

Chapter 10

Conclusion and Future Work

10.1 Conclusion

Information and Communication Technology (ICT) will play a critical role in the advancement of emerging economies by empowering people who live in developing regions through access to information, education and ability to self-organize. ***Connectivity*** is a central component in modern-day technology; however, current designs for wide-spread connectivity are extremely infrastructure-intensive. At the same time, emerging markets are often characterized with under-developed infrastructure, which coupled with low-income and sparse populations renders ubiquitous network deployments infeasible. This, in turn, is a major roadblock in access and adoption of technology by residents in developing countries.

This dissertation integrates the field of *wireless networks* with *Information and Communication Technology for Development (ICTD)* and focuses on design of *resource-aware modular wireless solutions* to connect communities in infrastructure-challenged environments. The presented research approach has three cornerstones: (i) *identification of problems* in access and adoption of technology, (ii) *design and implementation of network systems* to solve the identified problems, and (iii) *development of applica-*

tions that harness network infrastructure to improve the day-to-day life of residents in developing countries. Some of the fundamental problems in this domain are in core network systems design and implementation, while others are in studying specific environmental factors that influence network design, performance and utilization. The contributions of my research to date lie in the intersection of these problem sets and have the following primary dimensions:

1. **Understanding rural residents'** network usage patterns through analysis of large-scale mobile and fixed network traces generated by targeted communities [88, 119, 153–156].

2. **Design of network systems** optimized to the unique usage behavior of rural communities to deliver cellular and broadband Internet connectivity solutions in rural areas [120, 152, 153].

Our work demonstrates that there are multiple opportunities to rethink current designs and bring down the cost for rural network deployments, while accommodating immediate user needs. Our research has shown that rural communities exhibit high locality of communication needs. In line with this observation, we design resource and demand-aware *modular network systems* that can provide different tiers of connectivity depending on available resources for global access and user demand. The primary operation of such networks is to enable local communication within a community. Where outbound access is available, these networks can also deliver more global connectivity. Our system designs make use of various wide-range technologies spanning from plain GSM, to Dynamic Spectrum Access harnessing TV white spaces.

10.1.1 Understanding Rural Connectivity Demands

Realization of the utility of Internet and cellular access is the key ingredient to successful adoption. Poor network performance can easily impede the process of adoption by persistently failing to accommodate user demand and ultimately discouraging people from making use of connectivity. With this in mind, we evaluate performance and adoption of various connectivity technologies in developing rural regions and identify avenues for improvement.

Cellular networks. To understand cellular network usage, we conducted in-person interviews with users in rural Africa. We also analyzed large-scale highly-anonymized datasets by Orange Telecom from their network in Ivory Coast. Our interview results indicate that cellular access is of critical importance to the community and, while the reasons for adoption are not drastically different than those in Western users, the benefits for people in remote communities are much more pronounced. Furthermore, we learned that people rely on their cellphones to obtain locally-relevant information and poor reception have lead to many missed opportunities. Through such feedback we realize that local networking solutions have great potential for improved dissemination of information.

In our analysis of cellular network usage in Ivory Coast we were particularly interested in answering two questions. First, *how rural cellular network usage differs from that in urban areas* and second, *how do subscriber communities evolve in egocentric user networks*. We find out that urban users are much more globally-interconnected in their network usage, while rural residents tend to primarily call users within the same geographic location. These results are counter-intuitive to the current implementation of cellular networks and inform our design of network architectures that provide ubiquitous rural connectivity at a lower cost. Our community evolution analysis indicates

there is a potential for improved information dissemination through utilization of information relays in egocentric subscriber networks.

Contributions and impact: Our work is among the first to utilize sociological approaches such as interviews to inform localized communication system development. Further, we harness a carefully selected combination of cellular network traces and population density data to extract the unique characteristics of cellular network usage in rural areas.

Internet measurements. In our work on Internet measurements we collect a unique trace, capturing rural network usage. This trace, among other unique features, includes traffic before and after an eight-fold upgrade of the Internet gateway capacity in rural Macha, Zambia. We analyze *the implications of this eight-fold capacity increase on network performance and user experience* [156]. We show that the relationship between (i) network usage and performance and (ii) the available bandwidth is not trivial in the context of rural communities and is governed by the level of adoption of online services. Slow rural gateways typically support only basic web browsing and fall short in the face of more demanding services such as video streaming and large file upload/download. Thus, as more bandwidth is introduced, people attempt to harness this bandwidth to access a richer variety of services. However, if the available bandwidth is insufficient to meet the increasing demand associated with new service adoption, this can result in further deterioration of network performance and user experience.

Contributions and impact: We identify surprising aspects of Internet adoption that unveil novel research problems. Based on our analyses, we recommend a set of metrics that objectively capture the level of Internet adoption in rural areas [119]. Our findings are published in a highly-regarded conference in the field of ICTD [156]. Furthermore, our paper [119] is fundamental across disciplines in quantifying rural Internet adoption. Lastly, our work contributes a unique long-term network trace that allows analysis of Internet usage and performance in rural communities.

10.1.2 Connectivity for Infrastructure-Challenged Environments

We design and build connectivity solutions informed by our usage analysis, with the goal of widespread adoption. While the focus of our network studies and system deployments is in rural sub-Saharan Africa, our networking designs are well-suited for a broad spectrum of infrastructure-challenged environments. We design solutions both for cellular access to ensure rapid adoption, as well as more long-term solutions for wide-spread access through TV white spaces to provide high-speed local Internet connectivity in rural communities.

Cellular. To address the need for low-cost local cellular networks, we design *Kwiizya*, an open-source software and hardware system that leverages a generic IP backbone to provide voice connectivity and text messaging [153]. Kwiizya makes use of open-source implementation of GSM, which allows users to associate with their existing phones and SIM cards without any modifications. The GSM communication between users and the network is translated to VoIP by the base station, which allows Kwiizya to use generic IP backbone as opposed to expensive commercial-grade links. For authentication and switching services, we integrate several free open-source components that handle delivery of text and calls in the networks. Where outbound connectivity is available, Kwiizya can support calls and text messages to users outside of the village network. Finally, we enhance the existing open-source software with capability to support text message-based applications through instant message-to-text message functionality.

Contributions and impact: Our work on Kwiizya has received significant attention in the community. Our paper was presented at ACM MobiSys, a top wireless systems conference (acceptance rate 15.7%). In addition, our work was featured in multiple science and technology magazines, including the UK's New Scientist [67], and has reached readers in Africa through periodicals such as IT NEWS AFRICA [14]. We

have been contacted by health care organizations across Africa who have found that Kwiizya is well-aligned with their long-term vision for technological innovation. We are currently in discussion of new network deployments, which will enable further domain-specific research contributions. Kwiizya represents a realization of low-cost, localized infrastructure that improves everyday lives in rural areas. It also shows how inert regulations can hamper development. Kwiiziya requires flexible, localized GSM licensing, and since our work was published the discussion has started in the community about the revision of these licensing laws [67, 153].

Broadband. While mobile cellular networks often are the shortest path to wide-spread connectivity, *computer-based broadband* access is still necessary for rural residents to fully embrace the opportunities provided by the Internet. As we have shown through our network trace analyses, the limitations in network capabilities can have detrimental impact on user experience [156]. Thus we endeavour to design solutions for wide-spread wireless broadband in rural areas. Our focus is on TV white spaces, which with their favorable propagation properties, are a great candidate for wide-spread broadband network coverage. To this end we design *VillageLink*, a *combined PHY/MAC mechanism for efficient channel allocation*. We also design a spectrum sensing mechanism, TxMiner, that enables dynamic spectrum access and regulations beyond TV white spaces. VillageLink is designed to operate in a wide frequency band (50MHz to 800MHz) and makes use of existing TV antennas to connect end users. This design poses some unique research challenges in channel allocation due to wide-band operation and diverse antenna effects. In particular, the selection of transmission frequency can impact the existence of a link. To address this problem we make use of a two-step lightweight channel probing technique to understand the channel conditions at each communicating node. This creates a large solution space that calls for efficient methods to find the optimal channel allocation scheme. We design one such scheme

that guarantees real-time convergence and identifies the optimal channel allocation to minimize interference while maximizing throughput.

While VillageLink features a lightweight spectrum sensing technique, in a more dynamically-utilized environments and especially in bands beyond TV white spaces, real-time spectrum sensing performed by communicating parties becomes infeasible and might compromise the ability of the network agents to provide real-time services. To address this problem, there is a need for external spectrum sensing. To this end we design a spectrum sensing system called TxMiner that harnesses machine-learning techniques to determine spectrum occupancy and transmitter characteristics, thus aiding improved dynamic spectrum access.

Contributions and impact: Our work is the first to consider cost-efficient TV white spaces connectivity through utilization of existing TV antennas. We identify unique challenges related to channel allocation in a wide frequency band and across heterogeneous antenna types and address these challenges in a unified system design called *VillageLink* [120]. Our evaluation quantifies the importance of spectrum sensing for successful channel selection. We design the first spectrum-sensing technique that makes use of machine-learning techniques to understand spectrum occupancy.

10.2 Future Work

In the Western world, the problem of poor access performance is often solved through increase of available bandwidth. In rural developing regions, however, high prices, sparse populations and lack of guarantees for return of investments render this approach infeasible. The latter requires alternative solutions for improved online experience. To better understand the requirements for such alternative solutions, the introduction of a new technology in infrastructure-challenged regions needs to be carefully

planned and continually evaluated (Figure 10.1). This process starts with *studying the information needs* of targeted populations. Once this first stage is completed, a *new technology*, appropriate for the identified needs may be introduced. After a new technology is in place, a continuous *evaluation* of the performance and usage of this technology is necessary to assure smooth adoption. Any drawbacks identified in the process of evaluation should be addressed in an *improvement* phase, in which new features and technologies are developed.

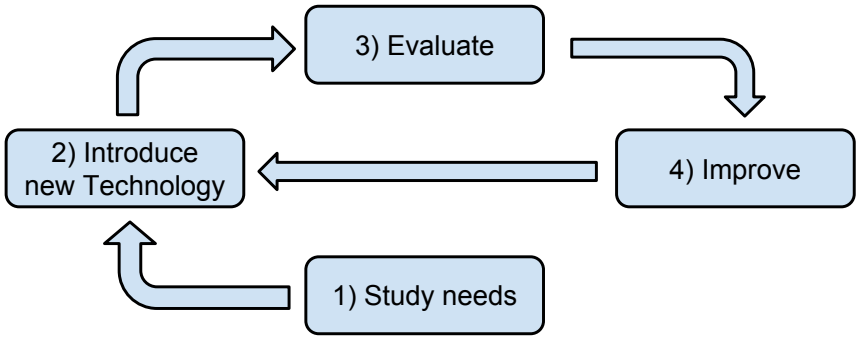


Figure 10.1: Cycle of new technology for infrastructure-challenged regions.

Future developments in the field of ICTD should strive to follow the methodology outlined above in order to develop network systems that provide high-quality user experience and allow equal participation of rural dwellers in Internet content access and generation. Such systems should not only allow access to modern services that is on par with that in the Western world but should also enable rural users to actively contribute local content. This dissertation presents several such solutions for rural cellular and Internet access. Moving forward towards equal rural connectivity, there are two important aspects to be considered: *continuous evaluation* of connectivity adoption and performance and the respective *improvement of existing systems*. The next two sections elaborate on these topics.

10.2.1 Continuous Evaluation of Connectivity Performance and Adoption

Current approaches for broadband measurement are quantitative, concentrated on estimating the number of users who have physical access to the Internet or the span of areas that are served by Internet service providers. The latter is often approximated through the number of providers serving a specific ZIP code [9]. Besides being quite coarse, these metrics do not capture the vast difference in the quality of access among users. While we do not argue for a different way of measuring the number of users, we claim that the quality of the Internet use experience among people in rural or remote communities requires a more elaborate approach that includes evaluation of both *network performance* and *adoption*.

10.2.1.1 Measuring performance

The first step towards understanding user satisfaction in Internet access is measuring network performance. In our work we have evaluated TCP performance to better understand if a network can meet usage demand. A first impression of the network state is given by the *offered load* as a fraction of the total gateway bandwidth; as the load approaches the bandwidth, network performance deteriorates. *Round-Trip Time* (RTT) is another indicator that reveals the networks ability to handle real-time applications: the larger the RTT the higher the probability of timeout, which leads to increased *packet losses* and *retransmissions*. Packet *payload size* describes well the ability of the network to handle the offered load. Evaluation of the *TCP control overhead* in a network is also indicative of performance: smaller amount of control packets indicates that the network is able to send larger portions of data at a time (i.e. has a larger *congestion window*), which means that the network is more efficient in transferring data without retransmissions. Separating TCP into *uplink flows* and *downlink flows* in com-

Table 10.1: Summary of proposed metrics, their units, and temporal interpretation.

Metric dimension	Units	Time-variant interpretation
Per-user bandwidth	Bits per second	Yes
Online affordances	Complex: a function of locally-relevant content, language, and supporting infrastructure	Yes
Location of access	Home/work/public terminal	No
Cost-benefit ratio	Complex: a function of the percent of per-person GNI and the economic impact of connectivity	No

bination with the above metrics allows evaluation of demand satisfaction. Particularly, one is able to discern if users are successfully able to access online services and upload their own content to the Internet. Lastly, in order to evaluate the effects of network performance on usage, one needs to continually evaluate the types of accessed online services.

10.2.1.2 Measuring adoption

The ultimate goal of a new technology is wide adoption. We argue that a comprehensive metric for measuring broadband adoption that includes all aspects of Internet access and usage requires a multidimensional approach. In Table 10.1 we summarize the dimensions that are, based on our analysis of Internet usage in rural Africa, the most important for efficient broadband access.

Internet connectivity is more than a question of physical access to the network. It is a matter of what the access can afford to the user that determines the utility of the Internet. Obstacles to efficient Internet utilization in rural Africa influence the way broadband adoption should be measured.

First, there is a clear impact of per-user Internet bandwidth that limits the type of applications that can be used. Rural areas are often deprived of access to high-speed fiber optic Internet backbone cables. Access is usually brought by satellite connectiv-

ity or other long-distance wireless links, which results in at least an order of magnitude slower connection than at-home broadband connectivity in the developed world. To offset the high cost of satellite access, the connection is shared among multiple users. In the villages we surveyed, hundreds of users share the same connection gateway. This results in an extremely low per-user bandwidth. ITU (2011) statistics, for example, show that the international Internet bandwidth in Africa reaches 937 bits per second per user, while in Europe it is 78,678 bits per second per user. Our research shows that bandwidth impacts the way files are shared and is one of the main obstacles to local content generation in rural Africa. Moreover, more Internet bandwidth is needed for an efficient online experience. The average Web page grew 36 times in just 15 years. Bandwidth-hungry applications such as video streaming services and online social networks are becoming more popular. Thus, per-user bandwidth must be considered with respect to the requirements of popular online applications at the time of the measurement.

The online realm is not of equal value to everyone. While a user in the urban developed world can engage in online commerce, obtain directions via online map services, and schedule a doctors appointment online, users in the rural developing world generally do not enjoy these benefits. Measuring the opportunities that one can obtain online requires a complex metric that takes into account the availability of online content that is relevant to the user. The relevance is connected to the language one speaks, the culture one is a part of, and the socio-economic and geographic context in which one resides. Measuring online affordances also must take into account the infrastructure that is not necessarily connected with Internet access, such as credit card banking and transport of goods, but still plays a role in the utility of some online services. Finally, online affordances should include the social aspect of using the Internet, as elaborated by the social affordances concept introduced by Wellman et al. [146].

Location of access critically impacts Internet usage. In our research we find that at-home connectivity leads to more leisurely Internet use with an emphasis on online social networking and content generation. Wyche et al. [149] show that public access leads to the so-called deliberate interaction model, where online activities must be pre-planned before the access actually happens. On the other hand, Best et al. [25] find that public access in African cyber-cafés has an educational note, and users engage in collaborative Web access to enhance their skills. Similarly, Rhinesmith [127] finds that Internet users in public libraries in Philadelphia, Pennsylvania, supplement, rather than replace, at-home access with public access. The latter is seen as a more social experience.

The cost of broadband puts Internet access into perspective with other basic necessities. Broadband connectivity requires up to a few hundred times proportionally higher investment from a user who lives in the developing world than from a user who lives in the developed world. Consequently, the benefits of the access have to be high. To obtain a clear picture of the effectiveness of broadband access, we propose inclusion of a cost-benefit analysis in the overall metric. Measuring the affordability of broadband can be done through identification of a percentage of per capita gross national income, as in the ITU (2011) study, for example. The benefit that Internet access brings is harder to measure, and we suggest investigation of the existing economic practices and online behavior to describe the impact of the Internet on ones living standard.

10.2.2 Future Rural Networks: Infrastructure-Limited Wireless Area Networks

Future research in networks for infrastructure-challenged environments should strive to develop reliable network systems and applications to enable full participation of rural residents in Internet content access and generation. While wireless networks have

lead the way in providing last hop access in more densely-populated areas, current network designs fall short in the case of rural developing regions. Current networks are (i) very infrastructure-intensive and (ii) rely on high bandwidth gateway access links to connect to the Internet. Unfortunately, these two resources are often unavailable in the rural developing context. To connect rural communities in infrastructure-deprived areas, I envision a new generation of *Infrastructure-Limited Wireless Area Networks (ILWAN)*. Such networks need to provide a plethora of services including plain voice, text messaging, and mobile and computer-based Internet. Thus, ILWANs will use a combination of local cellular solutions and wide-area TV white spaces networks to reach both mobile and stationary users. Such networks will make use of any available outbound resources, including basic commercial cellular networks (voice and text) and slow (satellite or otherwise) Internet gateways. Some unique challenges that arise in this context concern coexistence of commercial and local networks and efficient utilization of the existing resources by these local networks. More specifically, I envision the following avenues for fundamental research contributions.

1. **Detection** of existing resources. To leverage the existing network resources, local networks need to be cognitively-aware of the presence and operations of commercial networks. This requires elaborate spectrum sensing techniques that can differentiate between idle and occupied spectrum and can detect specific features of the ongoing transmissions.

2. **Coexistence** of commercial and local (Kwiizya-like) networks. This relationship can go two ways: local wireless networks can make use of the resources provided by commercial networks but can also enhance commercial networks operations through providing advanced services (e.g. mobile data).

3. **Limited resource operations.** As opposed to traditional wireless networks, where the bottleneck is in the last mile connectivity, in rural scenarios, with the advancement of last mile technologies, the bottleneck will shift onto the gateways. This will pose unique challenges in catering for full user experience that can be addressed in both networks as well as applications design.

Detection of available resources. A critical component of ILWANs that enables them to detect and utilize existing resources is their cognitive radio capability. A fundamental research problem is *how to recognize ongoing transmissions from raw spectrum measurements*. Traditionally, solving this problem has been a challenge due to the noisy nature of radio signals, which in turn makes it hard to detect ongoing transmissions. A new generation of sensing techniques can make use of machine-learning techniques to determine spectrum occupancy status. Ultimately, such techniques should be able to not only identify whether the spectrum is free or occupied, but more specifically how many transmitters are present, their bandwidths, whether they are mobile or static, their multiple access strategy (TDMA/FDMA) and direction. This methodology will provide a ***fundamental spectrum sensing approach***, which makes use solely of *raw spectrum measurements* and, as opposed to previous work, does not require prior knowledge of transmitter characteristics in order to identify transmissions. Our work on TxMiner makes a step in this direction. Future developments, however, will require enhanced identification of transmitter direction and mobility and should work with a plethora of signal propagation distributions to cater for transmitter detection in different wireless environments.

While this detection methodology has direct application in the design of ILWANs, it has broader applicability in (i) *improved DSA technology* through advanced spectrum sensing techniques, and (ii) in *spectrum policing and regulations*. Traditionally, practical Dynamic Spectrum Access (DSA) systems have focused on TV white spaces as

a medium. The latter have fairly static utilization in time and frequency because TV channels operate at a fixed bandwidth and either broadcast continuously or are silent. This makes the problem of scanning and frequency allocation fairly straightforward. To enable true opportunity for DSA technology, I consider DSA operations in frequency ranges beyond TV white spaces. In this context the problem of operation frequency selection becomes more complicated. Beyond TV white spaces, transmitters have much more dynamic characteristics, e.g. frequency-hopping (Bluetooth), aperiodicity (satellite transmissions) or varying channel width (802.11n/ac). To improve the state-of-the-art, DSA systems can utilize historical spectrum occupancy data when making decisions about operating spectrum. Furthermore, such historical data can help regulators to better understand geographic spectrum occupancy patterns and make informed decisions about spectrum policy to support creation of adequate spectrum regulations and bring forward DSA licensing efforts.

Coexistence of cellular networks. To provide reliable networks and services in rural communities, there is a need of a hybrid solution that allows coexistence of heterogeneous commercial and local cellular networks. From a user's perspective, the local and commercial cells will together be perceived as a single service. Such coexistence will allow plethora of applications, including (i) low-cost local communications, (ii) information services for local organizations (e.g. hospitals, schools, transportation firms), (iii) low-cost packet-switched data, and (iv) high-availability disaster-relief communications. To enable this coexistence, the local cell will need to acquire operations information from the commercial cell. The latter requires design of *cognitively-aware local cells* that utilize inexpensive Software-Defined Radios (SDR) to enable local GSM cells (such as our Kwiizya) to monitor and respond to the operations of other cells. This leads to unique challenges related to spectrum sensing, which can be

addressed through the functionality provided by our previously described methods for transmitter detection.

Current open-source implementations of the GSM stack support packet-switched services (such as GPRS) and are increasingly focused on bringing LTE services in the stack. Thus, low-cost local cellular networks solutions, can make use of generic IP gateways to connect to the Internet and can be used as mobile cellular platforms to *distribute Internet access to mobile users* in a remote community. Where commercial networks do not provide mobile data, this functionality can be used to enhance commercial networks' operations. Even where commercial data is available, the presence of local data-enabled cells provides new research opportunities for *mobile data offload* from commercial to local cellular networks.

Yet another challenge in this scope is *seamless user transition between multiple cellular networks*. This will require multiple networks to be used with the same SIM card. Since local cells allow open registration, users can opt in to register with their existing commercial network SIM cards. Then, association with either network can be triggered in one of two ways: either through (i) conscious choice by the user or through (ii) network-initiated association. While the former is straightforward, the latter poses interesting research challenges, which can be tackled through design of protocols that manipulate physical layer parameters of the local cellular network (e.g. transmission power and advertised neighbor frequencies) to trigger users to associate or disassociate from the local network.

Limited resource operations. While in commercial networks the last mile connectivity is typically the throughput bottleneck, in rural settings with slow Internet connectivity the bottleneck shifts to the gateways. This shift brings interesting research questions at the boundary between Internet gateways and local networks. The challenges of limited resource operations can be tackled both in network systems as well as ap-

plications design. To improve network systems design, one can make use of *real-time local caching techniques* and *novel mobile computing paradigms* such as *distributed clouds* and *cloudlets* to bring content and computational capability closer to the end user at the network edge. Subsequently, applications can harness these network technologies to deliver rich user experience by utilizing both locally-available content and computational power as well as global resources available on the Internet.

Such applications will have an immense potential for impact in many areas, such as health care (through access to remote databases, tele-medicine and information dissemination), education (through online or remote tutoring), transportation (through ad-hoc coordination of transport opportunities) and local farming, to name a few. Applications can harness low-cost Human-Computer Interaction (HCI) technologies, such as WiFi-enabled smartpens, to track paper-based health records or class work completion and offload tracked data to local cloudlets for processing and remote supervision. Ultimately, these applications will alleviate the load from ever overworked staff in remote institutions and bring outside experts' opinion in the day-to-day routines of local staff.

10.3 Summary

Access to technology will continue to drive development. Technologies that provide limited capacity are becoming available in infrastructure-challenged environments and while those limited capacities are a step towards bridging the digital divide what they create is a new form of *divide in the utility of Internet access*. Users in infrastructure-challenged environments are now limited by their available capacities in the types of services they can adopt. Such limitations will continue to take their toll on the economic, governmental and technological development of these communities. As a result there will always be a need for alternative solutions that bring rural users' technolog-

ical experience closer to that of the Western world. This dissertation advances the field by providing several such technologies that provide reliable low-cost connectivity in infrastructure-challenged areas. Multiple factors need to be considered in future designs, including appropriate interfaces, integration with existing systems and challenges with local governance. Thus, future developments will require not only further advances in networking systems, but also inter-field solutions both across computer science disciplines (on the edge of mobile systems, cloud computing and HCI), as well as beyond sciences and into humanities and economics.

Bibliography

[1] African Undersea Cables, http://manypossibilities.net/african-undersea-cables/.

[2] Dr Math – remote math tutoring using MXIT in South Africa. http://www.elearning-africa.com/eLA_Newsportal/mixing-it-with-dr-math-mobile-tutoring-on-demand/.

[3] European Commission urban-rural typology. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Urban-rural_typology. Accessed: 09/02/2013.

[4] FCC Notice of Inquiry, ET Docket No. 10-237, 25 FCC Rcd 16632 (2010).

[5] GeoIP Products, MaxMind, http://dev.maxmind.com/geoip/.

[6] ITU, Estimates of Ie and Bpl parameters for a range of CODEC types.

[7] Private communication with spectrum regulators from Philippines and Kenya.

[8] Rural Population in Cote D'Ivoire. http://www.tradingeconomics.com/cote-d-ivoire/rural-population-wb-data.html. Accessed: 03/02/2013.

[9] U.S. Government Accountability Office. Broadband deployment is extensive throughout the United States, but it is difficult to assess the extent of deployment gaps in rural areas. *Report to Congressional Committees*, 2006, http://www.gao.gov/new.items/d06426.pdf.

[10] *ITU-T Recommendation G.107, "'The E-Model, a computational model for use in transmission planning"*, December 1998.

[11] ITU-T Recommendation G.114, One-way transmission time. 2003.

[12] *ITU World Telecommunication/ICT Indicators Database*, 2013.

[13] LA building's lights interfere with cellular network, FCC says, February 2014.

[14] Experimental network connects Zambias most remote. *IT News Africa*, July, 2013, `http://www.itnewsafrica.com/2013/07/experimental-network-connects-zambias-most-remote/`.

[15] I. Abdeljaouad, H. Rachidi, S. Fernandes, and A. Karmouch. Performance analysis of modern TCP variants: A comparison of Cubic, Compound and New Reno. In *QBSC*, Kingston, ON, Canada, May, 2010.

[16] R. Ahmed and G. Karypis. Algorithms for mining the evolution of conserved relational states in dynamic networks. In *ICDM*, 2011.

[17] J. C. Aker and I. M. Mbiti. Mobile Phones and Economic Development in Africa. In *Journal of Economic Perspectives*, volume 24, 2010.

[18] A. Akue-Kpakpo. Study on international Internet connectivity in sub-Saharan Africa. *ITU-D*, March 2013.

[19] I. Allagui and J. Kuebler. The Arab Spring and the Role of ICTs. In *International Journal of Communication, Vol. 5*, pages 1435–1442, 2011.

[20] A. Anand, V. Pejovic, E. M. Belding, and D. L. Johnson. VillageCell: Cost effective cellular connectivity in rural areas. ICTD, Atlanta, Georgia, March, 2012.

[21] R. Anderson, E. Blantz, D. Lubinski, E. O'Rourke, M. Summer, and K. Yousoufian. Smart connect: last mile data connectivity for rural health facilities. In *NSDR*, San Francisco, CA, 2010.

[22] C. R. Andrew Blake, Pushmeet Kohli. Mit press. 2011.

[23] Y. Anokwa, C. Dixon, G. Borriello, and T. Parikh. Optimizing high latency links in the developing world. In *WiNS-DR*, San Francisco, CA, September, 2008.

[24] L. Antova, T. Jansen, C. Koch, and D. Olteanu. Fast and simple relational processing of uncertain data. In *ICDE*, April 2008.

[25] M. Best, B. Kollyani, and S. Garg. Sharing in public: Working with others in Ghanian cybercafes. *International Conference on Information and Communication Technologies and Development (ICTD 2012)*, March 2012.

[26] V. Blondel, G. Krings, and I. Thomas. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. In *The e-journal for academic research on Brussels, Issue 42*, October, 2010.

[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[28] J. Blumenstock and N. Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In *ICTD*, 2010.

[29] P. Bogdanov, M. Mongiovi, and A. Singh. Mining heavy subgraphs in time-evolving networks. In *ICDM*, 2011.

[30] P. Bremaud, editor. *Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues*. Springer, 1999.

[31] E. Brewer, M. Demmer, M. Ho, R. Honicky, J. Pal, M. Plauche, and S. Surana. The Challenges of Technology Research for Developing Regions. *Pervasive Computing*, 2006.

[32] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 665–674, New York, NY, USA, 2011. ACM.

[33] F. Calabrese and C. Ratti. Real Time Rome. In *Networks and Communication Studies – Official journal of the IGU's Geography of Information Society Commission, 20:3, 247-258*, 2006.

[34] F. Calabrese, Z. Smoreda, V. Blondel, and C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. In *PLoS ONE 6(7)*, 2010.

[35] B. Campbell, J. Rosenberg, H. Schulzrinne, C. Huitema, and D. Gurle. Internet Engineering Task Force RFC3428: Session Initiation Protocol (SIP) extension for Instant Messaging. December 2002.

[36] R. Cavallo and M. Pittarelli. The theory of probabilistic databases. In *VLDB*, 1987.

[37] R. Chaudhri, G. Borriello, and W. Thies. FoneAstra: making mobile phones smarter. In *NSDR*, San Francisco, CA, 2010.

[38] K. Chebrolu, B. Raman, and S. Sen. Long-distance 802.11b links: Performance Measurements and Experience. MobiCom '06, Los Angeles, CA, USA, 2006.

[39] D. Chen, S. Yin, Q. Zhang, M. Liu, and S. Li. Mining spectrum usage data: A large-scale spectrum measurement study. MobiCom '09, Beijing, China, 2009.

[40] J. Chen, L. Subramanian, and E. Brewer. SMS-based web search for low-end mobile devices. In *MOBICOM*, Chicago, IL, 2010.

[41] `http://www.childcount.org/`.

[42] A. Chowdhery, R. Chandra, P. Garnett, and P. Mitchell. Characterizing Spectrum Goodness for Dynamic Spectrum Access. 50th Allerton Conference on Communication, Control and Computing, Monticello, Illinois, October, 2012.

[43] C. Cordeiro, K. Challapali, D. Birru, B. Manor, and S. Diego. IEEE 802.22: An Introduction to the First Wireless Standard based on Cognitive Radios. *Journal of Communications*, 1(1):38–47, 2006.

[44] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal*, 2007.

[45] M. de Bruijn, F. B. Nyamnjoh, and I. Brinkman. Mobile phones: The New Talking Drums of Everyday Africa. 2009.

[46] N. L. Dell, S. Venkatachalam, D. Stevens, P. Yager, and G. Borriello. Towards a point-of-care diagnostic system: automated analysis of immunoassay test data on a cell phone. In *NSDR*, Bethesda, MD, 2011.

[47] B. DeRenzi, G. Borriello, J. Jackson, V. S. Kumar, T. S. Parikh, P. Virk, and N. Lesh. Mobile phone tools for field-based health care workers in low-income countries. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 78(3):406–418, 2011.

[48] B. Derenzi, T. Parikh, M. Mitchell, M. Chemba, D. Schellenberg, N. Lesh, C. Sims, W. Maokola, Y. Hamisi, and G. Borriello. e-IMCI: Improving Pediatric Health Care in Low-Income Countries. In *CHI*, 2008.

[49] J. Donner. The rules of beeping: exchanging messages using missed calls on mobile phones in sub-Saharan Africa. *International Communications Association*, 2005.

[50] B. Du, M. Demmer, and E. Brewer. Analysis of WWW traffic in Cambodia and Ghana. In *WWW*, Edinburgh, Scotland, May 2006.

[51] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. John Wiley and Sons, Inc., 2001.

[52] N. Eagle, Y. de Montjoye, and L. Bettencourt. Community computing: Comparisons between rural and urban societies using mobile phone data. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, pages 144–150, 2009.

[53] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. In *Science 21 May 2010: Vol. 328 no. 5981 pp. 1029-1031*, May 2010.

[54] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. In *PNAS, Vol. 106, No. 36.*, 2009.

[55] H. S. F. Fraser and S. J. D. McGrath. Information technology and telemedicine in sub-Saharan Africa. *BMJ*, 321:465–466, August 2000.

[56] S. Gabrysch, V. Simushi, and O. M. Campbell. Availability and distribution of, and geographic access to emergency obstetric care in Zambia. *International Journal of Gynecology and Obstetrics*, 114(2):174 – 179, 2011.

[57] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[58] A. Goldsmith. *Wireless communications*. Cambridge Univ Pr, 2005.

[59] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, Jun 2008.

[60] S. Ha, I. Rhee, and L. Xu. CUBIC: a new TCP-friendly high-speed TCP variant. *SIGOPS Oper. Syst. Rev.*, 42:64–74, July 2008.

[61] S. Habeenzu. Zambia ICT Sector Performance Review 2009/2010. *Research ICT Africa*, 2010.

[62] B. Hajek. Cooling Schedules for Optimal Annealing. *Mathematics of Operations Research*, 13:311–319, 1988.

[63] K. Harrison, S. M. Mishra, and A. Saha. How Much White-Space Capacity Is There? In *DySpan'10*, Singapore, April 2010.

[64] K. Heimerl, K. Ali, J. Blumenstock, B. Gawalt, and E. Brewer. Expanding Rural Cellular Networks with Virtual Coverage. In *NSDI*, 2013.

[65] K. Heimerl and E. Brewer. The Village Base Station. In *NSDR'10*, San Francisco, CA, June 2010.

[66] K. Heimerl, S. Hasan, K. Ali, E. Brewer, and T. Parikh. Local, sustainable, small-scale cellular networks. ICTD '13, Cape Town, South Africa, 2013.

[67] H. Hodson. Wi-fi-hopping brings phone signal to remote villages. *New Scientist*, June, 2013, `http://www.newscientist.com/article/mg21829235.900-wifihopping-brings-phone-signal-to-remote-villages.html`.

[68] S. S. Hong and S. R. Katti. Dof: a local wireless information plane. SIGCOMM, 2011.

[69] S. Ihm, K. Park, and V. S. Pai. Towards understanding developing world traffic. In *NSDR*, San Francisco, CA, June, 2010.

[70] T. Imieliński and W. Lipski, Jr. Incomplete information in relational databases. *J. ACM*, 31(4):761–791, Sept. 1984.

[71] International Telecommunications Union. The World in 2011; Facts and Figures. http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf, 2011.

[72] International Telecommunications Union. The World in 2013; Facts and Figures. `http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf`, 2013.

[73] S. Isaacman and M. Martonosi. Low-infrastructure methods to improve internet access for mobile users in emerging regions. WWW '11, Hyderabad, India, 2011.

[74] R. Jain, D. Chiu, and W. Hawe. A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems. Technical report, DEC Research Report TR-301, 1984.

[75] D. L. Johnson, E. M. Belding, K. Almeroth, and G. van Stam. Internet usage and performance analysis of a rural wireless network in Macha, Zambia. In *NSDR*, San Francisco, CA, June, 2010.

[76] D. L. Johnson, E. M. Belding, and G. van Stam. Network traffic locality in a rural african village. ICTD '12, Atlanta, Georgia, 2012.

[77] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. VillageShare: Facilitating content generation and sharing in rural networks. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12, pages 7:1–7:10, New York, NY, USA, 2012. ACM.

[78] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. Traffic characterization and internet usage in rural Africa. In *WWW*, Hyderabad, India, March, 2011.

[79] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. VillageShare: Facilitating content generation and sharing in rural networks. In *ACM DEV*, Atlanta, GA, March, 2012.

[80] D. L. Johnson and K. Roux. Building rural wireless networks: lessons learnt and future directions. WiNS-DR '08, San Francisco, California, USA, 2008.

[81] J. G. Jørn Klungsøyr. Using mobile phones to track immunizations. *Wireless Health Organization*, 2010.

[82] M. Kam, D. Ramachandran, V. Devanathan, A. Tewari, and J. Canny. Localized iterative design for language learning in underdeveloped regions: the pace framework. In *CHI*, San Jose, CA, 2007.

[83] B. Kauffmann, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki, and C. Diot. Measurement-Based Self Organization of Interfering 802.11 Wireless Access Networks. In *INFOCOM'07*, Anchorage, AK, May 2007.

[84] M.-S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. *Proc. VLDB Endow.*, 2(1):622–633, Aug. 2009.

[85] A. B. King. Web site optimization, `http://www.websiteoptimization.com/speed/tweak/average-web-page/`.

[86] S. Kirkpatrick, C. D. G. Jr., and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220:671–680, 1983.

[87] V. Kone, L. Yang, X. Yang, B. Y. Zhao, and H. Zheng. On the feasbility of effective opportunistic spectrum access. IMC 2010.

[88] V. Kone, M. Zheleva, M. Wittie, B. Y. Zhao, E. M. Belding, H. Zheng, and K. Almeroth. AirLab: Consistency, Fidelity and Privacy in Wireless Measurements. *SIGCOMM Comput. Commun. Rev.*, 41(1):60–65, Jan. 2011.

[89] G. Krings, M. Karsai, S. Bernhardsson, V. Blondel, and J. Saramaki. Effects of time window size and placement on the structure of an aggregated communication network. In *EPJ Data Science*, May 2012.

[90] A. Kumar, J. Chen, M. Paik, and L. Subramanian. ELMR: Efficient Lightweight Mobile Records. In *MobiHeld*, Barcelona, Spain, 2009.

[91] D. Li and J. Gross. Distributed TV spectrum allocation for cognitive cellular network under game theoretical framework. In *DySPAN'12*, 2012, address = Bellevue, WA, month = October, owner = mariya, timestamp = 2013.08.07.

[92] M. Ma and D. H. Tsang. Joint Design of Spectrum Sharing and Routing with Channel Heterogeneity in Cognitive Radio Networks. *Physical Communication*, 2(1-2):127–137, March 2009.

[93] MachaWorks. `http://www.machaworks.org/`.

[94] V. Mancuso and S. Alouf. Reducing costs and pollution in cellular networks. *Communications Magazine, IEEE*, 49(8):63–71, 2011.

[95] M. Materu-Behitsa and B. D. Diyamett. Tanzania ICT Sector Performance Review 2009/2010. *Research ICT Africa*, 2010.

[96] K. W. Matthee, G. Mweemba, A. V. Pais, G. van Stam, and M. Rijken. Bringing Internet connectivity to rural Zambia using a collaborative approach. In *ICTD*, Bangalore, India, 2007.

[97] V. W. A. Mbarika. Is telemedicine the panacea for Sub-Saharan Africa's medical nightmare? *Commun. ACM*, 47(7):21–24, July 2004.

[98] I. Mbiti and D. N. Weil. Mobile Banking: The Impact of M-Pesa in Kenya. Working Paper 17129, National Bureau of Economic Research, June 2011.

[99] M. A. McHenry, P. A. Tenhula, D. McCloskey, D. A. Roberson, and C. S. Hood. Chicago Spectrum Occupancy Measurements and Analysis and a Long-term Studies Proposal, Aug 2006.

[100] V. Mhatre, K. Papagiannaki, and F. Baccelli. Interfernece Mitigation through Power Control in High Density 802.11 WLANs. In *INFOCOM'07*, Anchorage, AK, May 2007.

[101] Microsoft Corporation. Microsoft Windows Server 2003 TCP/IP Implementation Details. `http://www.microsoft.com/en-us/download/details.aspx?id=13902`, 2007.

[102] M. Minges. Mobile cellular communications in the Southern African region. *Telecommunications Policy*, 23(7):585–593, 1999.

[103] A. Mishra, S. Banjaree, and W. Arbaugh. Weighted Coloring Based Channel Assignment for WLANs. *Mobile Comput. Commun. Rev.*, 9:19–31, 2005.

[104] T. Moscibroda, R. Wattenhofer, and Y. Weber. Protocol Design Beyond Graph-Based Models. In *HotNets'06*, 2006.

[105] J. Mpala and G. van Stam. OpenBTS, a GSM experiment in rural Zambia. Africomm, Yaounde, Cameroon, 2012.

[106] R. Murty, R. Chandra, T. Moscibroda, and P. V. Bahl. Senseless: A database-driven white spaces network. *IEEE Transactions on Mobile Computing*, 11(2):189–203, 2012.

[107] V. Ndou. E-government for developing countries: opportunities and challenges. In *The Electronic Journal of Information Systems in Developing Countries, Vol 18*, pages 1–24, 2004.

[108] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[109] G. Nychis, T. Hottelier, Z. Yang, S. Seshan, and P. Steenkiste. Enabling MAC Protocol Implementations on Software-Defined Radios. In *NSDI'09*, Boston, MA, April 2009.

[110] Y. B. Okwaraji, S. Cousens, Y. Berhane, K. Mulholland, and K. Edmond. Effect of geographical access to health facilities on child mortality in rural ethiopia: A community based cross sectional study. *PLoS ONE*, 7, 03 2012.

[111] B. G. Omwenga and P. W. Githinji. Buzzenger: Asynchronous (Time-Sliced) Missed Call Duration Messaging. In *NSDR*, Boston, MA, 2012.

[112] J. Onnela, S. Arbesman, M. Gonzalez, A. Barabasi, and N. Christakis. Geographic constraints on social network groups. *PLoS ONE*, 6(4):e16939, 04 2011.

[113] `http://www.openxdata.org/`.

[114] M. Paik, J. Chen, and L. Subramanian. Epothecary: cost-effective drug pedigree tracking and authentication using mobile phones. In *MobiHeld*, Barcelona, Spain, 2009.

[115] T. Parikh, N. Patel, and Y. Schwartzman. A survey of information systems reaching small producers in global agricultural value chains. In *ICTD*, Bangalore, India, December 2007.

[116] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural India. In *CHI*, Atlanta, GA, 2010.

[117] N. Patel, S. Klemmer, and T. Parikh. An asymmetric communications platform for knowledge sharing using cheap mobile phones. In *ACM Symposium on User Interface Software and Technology (UIST)*, Santa Barbara, CA, October 2011.

[118] R. Patra, S. Nedevschi, S. Surana, A. Sheth, L. Subramanian, and E. Brewer. WiLDNet: Design and Implementation of High Performance WiFi Based Long Distance Networks. In *NSDI'07*, Cambridge, MA April 2007.

[119] V. Pejovic, D. Johnson, M. Zheleva, E. Belding, L. Parks, and G. van Stam. The Bandwidth Divide: Obstacles to efficient broadband adoption in rural Sub-Saharan Africa. *International Journal of Communication (IJoC)*, 6 (2012).

[120] V. Pejovic, D. Johnson, M. Zheleva, A. Lysko, and E. Belding. VillageLink: Wide-Area Wireless Coverage. COMSNETS'14, Bangalore, India, January, 2014.

[121] A. Pentland, R. Fletcher, and A. Hasson. Daknet: rethinking connectivity in developing nations. *Computer*, 37(1):78–83, 2004.

[122] D. Ramachandran, J. Canny, P. D. Das, and E. Cutrell. Mobile-izing health workers in rural India. In *CHI*, Atlanta, GA, 2010.

[123] K. N. Ramachandran, E. M. Belding, K. C. Almeroth, and M. M. Buddhikot. Interference-Aware Channel Assignment in Multi-Radio Wireless Mesh Networks. In *INFOCOM'06*, Barcelona, Spain, April 2006.

[124] B. Raman and K. Chebrolu. Design and Evaluation of a new MAC Protocol for Long-Distance 802.11 Mesh Networks. In *MobiCom'05*, Cologne, Germany, August/September 2005.

[125] S. Rayanchu, A. Patro, and S. Banerjee. Airshark: Detecting non-wifi rf devices using commodity wifi hardware. IMC '11, Berlin, Germany, 2011.

[126] A. Reda, S. Panjwani, and E. Cutrell. Hyke: a low-cost remote attendance tracking system for developing regions. In *NSDR*, Washington, D.C., 2011.

[127] C. Rhinesmith. Free library hot spots: Redefining broadband adoption in Philadelphias low-income communities. *Defining and Measuring Meaningful Broadband Adoption Workshop.*, Aptil 2012.

[128] Rhizomatica. `http://rhizomatica.org/`.

[129] C. C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[130] J. Rosenberg, H. Schulzrinne, C. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. Internet Engineering Task Force RFC3261: SIP: Session Initiation Protocol. June 2002.

[131] A. Sheth, S. Nedevschi, R. Patra, S. Surana, E. Brewer, and L. Subramanian. Packet Loss Characterization in WiFi-Based Long Distance Networks. INFOCOM 2007, Anchorage, AK, USA, 2007.

[132] A. S. Sife, E. Lwoga, and C. Sanga. New technologies for teaching and learning: Challenges for higher learning institutions in developing countries. In *International Journal of Education and Development using ICT, Vol. 3, Issue 2*, pages 57–67, 2007.

[133] R. Sinha, C. Papadopoulos, and J. Heidemann. Internet packet size distributions: Some observations. Technical Report ISI-TR-2007-643, USC/Information Sciences Institute, May 2007. Orignally released October 2005 as web page `http://netweb.usc.edu/~rsinha/pkt-sizes/`.

[134] I. Stojmenovic, editor. *Handbook of Wireless Networks and Mobile Computing*. Wiley-Interscience, 1st edition, February 2002.

[135] D. Suciu, D. Olteanu, C. Ré, and C. Koch. Probabilistic databases. *Synthesis Lectures on Data Management*, 3(2):1–180, 2011.

[136] S. Surana, R. Patra, S. Nedevschi, and E. Brewer. Deploying a rural wireless telemedicine system: Experiences in sustainability. *Computer*, 41(6):48 –56, June 2008.

[137] S. Surana, R. Patra, S. Nedevschi, M. Ramos, L. Subramanian, Y. Ben-David, and E. Brewer. Beyond pilots: keeping rural wireless networks alive. In *NSDI*, San Francisco, CA, April 2008.

[138] L. Sydell. FCC Eyes Broadband For Indian Reservations. online, June 2010. http://www.npr.org/templates/story/story.php?storyId=128004928.

[139] P.-N. Tan, M. Steinbach, and V. Kumar, editors. *Introduction to Data Mining*. Addison Wesley, 2005.

[140] Default TTL Values in TCP/IP, `http://www.map.meteoswiss.ch/map-doc/ftp-probleme.htm`.

[141] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3:71–86, January 1991.

[142] P. van Hoorik and F. Mweetwa. Use of Internet in rural areas of Zambia. In *IST Africa*, Windhoek, Namibia, May 2008.

[143] L. Vannini and H. le Crosnier. *NET.LANG: Towards the multilingual cyberspace*. C&F edition, March 2012.

[144] W. W. Vithanage and A. S. Atukorale. Bassa: a time shifted web caching system for developing regions. NSDR '11, Bethesda, Maryland, USA, 2011.

[145] M. Wellens and P. Mahonen. Lessons learned from an extensive spectrum occupancy measurement campaign and a stochastic duty cycle model. TRIDENTCOM, april 2009.

[146] B. Wellman, A. Quan-Haase, J. Boase, W. Chen, K. Hampton, and I. D. K. Miyata. The social affordances of the Internet for networked individualism. *Journal of Computer-Mediated Communication*, 2006.

[147] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee4. Quantifying the impact of human mobility on malaria. In *Science, Vol. 338, No. 6104. (12 October 2012), pp. 267-270*, October 2012.

[148] S. P. Wyche, T. N. Smyth, M. Chetty, P. M. Aoki, and R. E. Grinter. Deliberate interactions: characterizing technology use in Nairobi, Kenya. CHI '10, Atlanta, Georgia, USA, 2010. ACM.

[149] S. P. Wyche, T. N. Smyth, M. Chetty, P. M. Aoki, and R. E. Grinter. Deliberate Interactions: Characterizing Technology Use in Nairobi, Kenya. In *CHI'10*, Atlanta, GA, April 2010.

[150] L. Yang, W. Hou, L. Cao, B. Y. Zhao, and H. Zheng. Supporting Demanding Wireless Applications with Frequency-Agile Radios. NSDI'10, San Jose, California, 2010.

[151] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Transactions on Information Theory 51 (7), page 22822312, July 2005.

[152] M. Zheleva, A. Chowdhery, R. Chandra, A. Kapoor, P. Garnett, P. Mitchell, and E. Belding. TxMiner: Identifying Transmitters in Real World Spectrum Measurements. in Submission.

[153] M. Zheleva, A. Paul, D. L. Johnson, and E. Belding. Kwiizya: Local Cellular Network Services in Remote Areas. MobiSys'13, Taipei, Taiwan, 2013.

[154] M. Zheleva, P. Schmitt, M. Vigil, and E. Belding. Bringing Visibility to Rural Users in Côte D'Ivoire. ICTD'13, Cape Town, South Africa, December, 2013.

[155] M. Zheleva, P. Schmitt, M. Vigil, and E. Belding. Community Detection in Cellular Network Traces. ICTD'13, Cape Town, South Africa, December, 2013.

[156] M. Zheleva, P. Schmitt, M. Vigil, and E. Belding. The Increased Bandwidth Fallacy: Performance and Usage in Rural Zambia. ACM DEV'13, Cape Town, South Africa, December, 2013.